VISIT…

LANZAROTE
Caliente.COM

语言与计算机丛书

语料库语言学

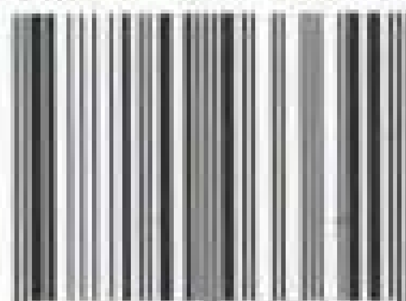
黄昌宁 李涓子 著



商务印书馆

语言与计算机丛书

语料库语言学



ISBN 7-100-03364-0/H · 850
定价：14.00 元

语言与计算机丛书

语料库语言学

黄昌宁　李涓子　著



商务印书馆
2002年·北京

图书在版编目(CIP)数据

语料库语言学/黄昌宁,李涓子著.—北京:商务印书馆,2002
(语言与计算机丛书)

ISBN 7-100-03364-0

Ⅰ.语… Ⅱ.①黄…②李… Ⅲ.计算机应用-语言学-研究 Ⅳ.H0

中国版本图书馆CIP数据核字(2001)第057971号

所有权利保留。
未经许可,不得以任何方式使用。

语言与计算机丛书
YŬLIÀOKÙ YŬYÁNXUÉ
语料库语言学
黄昌宁 李涓子 著

商务印书馆出版
(北京王府井大街36号 邮政编码100710)
商务印书馆发行
北京民族印刷厂印刷
ISBN 7-100-03364-0/H·850

2002年4月第1版 开本 787×1092 1/32
2002年4月北京第1次印刷 印张 9⅜

定价:14.00元

目　录

第1章　绪论 …………………………………………… (1)

第一节　什么是语料库? 什么是语料库语言学? …… (1)

第二节　语料库语言学的发展历史 ……………………… (3)

第三节　语料库语言学的发展方向及前景 ………… (13)

第四节　计算机在语料库语言学中的作用 ………… (15)

第五节　语料库语言学的研究内容 ……………… (17)

第六节　本书的编排 …………………………………… (21)

第2章　语料库的设计与开发 ……………………… (22)

第一节　语料库设计和编纂中的问题 …………… (22)

第二节　建设一个语料库 ………………………… (35)

第三节　语料库的类型 …………………………… (43)

第四节　国外语料库介绍 ………………………… (44)

第五节　汉语语料库的建设 ……………………… (68)

第3章　语料库的加工和管理技术 …………… (101)

第一节　语料的索引及其应用 ………………… (101)

第二节 语料库语言学中的统计…………………… (106)
第三节 逐词索引软件及其应用…………………… (120)
第四节 语料库标注……………………………… (139)

第4章 基于语料库方法的语言学研究 ……… (153)
第一节 语言研究中的语料库方法……………… (153)
第二节 现代汉语句型统计与研究……………… (161)
第三节 词典学研究……………………………… (167)
第四节 汉语名词的语义分类研究……………… (200)
第五节 词汇—语法问题调查…………………… (207)
第六节 语域变体(register variation)研究 ……… (212)

第5章 语料库方法在计算语言学中的应用
………………………………………… (220)
第一节 汉语文本中交集型切分歧义的研究……… (220)
第二节 汉语基本名词短语识别研究…………… (244)
第三节 基于结构词义空间的汉语词义排歧模型
………………………………………… (261)
附录 词性标记集 ……………………………… (278)

参考文献 ……………………………………… (280)

第1章 绪 论

"语料库语言学已经成为语言研究的主流。基于语料库的研究不再是计算机专家的独有领域,它正在对语言研究的许多领域产生愈来愈大的影响"。这是汤姆斯(Thomas)等人1996年为祝贺语料库语言学的主要奠基人和倡导者里奇(Leech)六十诞辰而编纂的语料库语言学研究论文集的开场白。近年来,对语料库语言学类似的说法频频见于导论和方法论的专著及教科书中,它不仅仅是语料库语言学家的自誉,而且正在成为整个语言学界的共识[1]。

第一节 什么是语料库?什么是语料库语言学?

语料库(corpus)顾名思义就是存放语言材料的仓库(或

数据库)。传统上,语言学家用语料库这个术语表示可作为语言研究基础的、大量自然出现的语言数据。这些语料库可以由书面语和(或)口语的样本组成,并通常被用来代表一种特定的语言或语言变体。在计算机出现之前,研究者——特别是词典编纂者,也有语料库,只是规模小、范围窄,因而难以在学术界形成气候。近 40 年以来,语料库这个术语通常指以电子形式保存的语言材料,并被广泛用于语言研究和语言工程。随着计算机功效的成倍增长,语料库在规模、多样性和使用方便等方面都发生了剧烈的变化。与此同时,为了存取和加工语料库所拥有的信息,已经开发了大量专用的软件。计算机语料库迅速成为语言研究的一种普遍资源,现在世界上已经建立了许多规模较大的语料库,有些是国家级的,有些是大学和词典出版商联合研制的。另外,由于个人电脑的迅猛发展,存储数据的硬磁盘造价持续下降,研究者个人也开始建立适合自己研究兴趣的小型语料库。

虽然语料库语言学研究已经历了不短的历史,但还没有一个公认的定义。下面引述两个见诸书本的定义:

定义 1:以现实生活中人们运用语言的实例为基础进行的语言研究,称为语料库语言学。(McEnery & Wilson, 1996)[2]

定义 2:以语料为语言描写的起点,或以语料为验证有关语言假说的方法,称为语料库语言学。(Crystal, 1991)[3]

从上述两个定义可见,作为一个学科的名称"语料库语言

学"与"语法学"或"语义学"不同,它不属于语言自身某个侧面的研究,而是一种以语料库为基础的语言研究方法。它实际上包括两方面的内容:一是对自然语料进行加工、标注,二是用已经标注好的语料进行语言研究和应用开发。

第二节 语料库语言学的发展历史

语料库语言学作为一种语言研究的方法,可以追溯到上个世纪,甚至更为久远。文献[1]对语料库语言学进行了论述,在此现在一般以乔姆斯基(N. Chomsky)转换生成语法的兴衰史为参照点,将语料库语言学的发展历史分为如下三个时期[1]。

一、早期的语料库语言学

早期语料库语言学是指20世纪50年代中期以前,即以乔姆斯基提出转换生成语法理论之前的所有基于语言材料的语言研究。在50年代,语料库在语言研究中曾被广泛使用,主要集中体现在以下几个方面。

1. 语言习得

语言习得是较早普遍用语料为研究方法的一个领域。19世纪70年代,在欧洲兴起了儿童语言习得研究的第一个高

潮。当时许多研究素材来自父母对其子女话语发展的观察日记。据悉,这些日记作为原始资料,不仅是当时 Preyer[4]和 Stern[5]等人提出理论假说的依据,而且时至今日仍是许多学者的研究材料之一。自本世纪 30 年代以来,语言学家和心理语言学家提出了许多关于儿童在不同年龄段的语言发展模式,这些模式大都建立在对儿童自然话语的大量观察材料的基础上。

2. 方言学

方言学从其产生以来,就与语料结下了不解之缘。在西方,方言学脱胎于 19 世纪的历史比较语言学,最初的兴趣主要是,运用直接法所获得的有关单音不同分布的事实来绘制方言地图。"方言研究者手持笔记本,后来是手提录音机,记下或录下他所遇到的一切方言材料。此种采样方式至今仍为某些业余研究者所沿用,它对于研究方言词汇的分布有一定价值"(Francis,1980)[6]。在我国,运用语料的研究方法可追溯至周秦。据南朝应劭《风俗同义序》"周、秦常以岁八月遣輶轩之使,求异代方言。"我国汉语方言学第一部著作《方言》就是这种方法的产物。据载,扬雄非常喜爱方言,他利用考廉(略等于后代的举人)和士兵集中在首都的方便,普遍地进行走访,不断积累材料,坚持编撰整理,经过 27 年的艰苦努力,终成《有輶轩使者绝代语释别国方言》。

3. 语言教学

Fries、Traver 和 Bonger (1947)是使用语料研究外语教学

法的语言学家。正如 Kennedy 所说(1992)[7],在 20 世纪的前 50 年中,语料库与外语教学有着密切的联系。外语教学中使用的词汇表,往往都是直接从语料中统计的。它对控制外语的学习过程十分重要。

4. 句法和语义

一些语言学家用语料库研究语言的描述。如:语言学家 Fries(1952)在语料库调查的基础上建立了英语的描写语法[8]。这项工作比 80 年代后期语言学家夸克(Quirk)等用语料库方法编撰的《英语语法大全》,早了 30 年。

5. 音系研究

利用自然语料开展音系研究,在西方当首推早期的结构主义语言学家,如 F. Boas 和 E. Sapir 等人。他们注重"野外工作",强调语料获取的自然性和语料分析的客观性。这些都为后来的语料库语言学所继承和发展。

20 世纪 50 年代中前期,在实证主义和行为主义思潮的影响下,总的来说在语言研究中占主导的是经验主义。这种气氛无疑促进了对语料的重视,使其成为当时的热点之一。特别是在美国,以哈里斯(Harris)等人为代表的后布龙菲尔德结构主义语言学家,视语料为语言学的唯一研究对象。在他们看来,直觉证据是第二位的,是靠不住的,应该扬弃。

二、乔姆斯基的转换生成语法时期

1957年乔姆斯基《句法理论》[9]及其以后一系列论著的发表，根本改变了语料库语言学的上述发展状况。语言学研究的主流方法也随之从经验主义(empiricism)转向理性主义(rationalism)。在这段时期中，笛卡儿的理性主义占据了主导地位，经验主义几乎无立足之地，被视为经验主义产物的各种语料库自然被完全否定。

理性主义研究方法认为，人的很大一部分语言知识是生来俱有的，是遗传决定的。而与理性主义相对峙的经验主义则认为，人的知识只是通过感官输入，并经过某些简单联想与一般化的操作而得到。人并非生来俱有一套有关语言的原则和处理方法。由于在语言学中乔姆斯基的内在语言或语言能力说被广泛接受，理性主义方法，从20世纪60年代到70年代长达20年的时间里，实际上主宰了欧美众多国家的语言学研究。

乔姆斯基及其转换生成语法学派批判早期语料库研究方法的主要论点如下：

1. 基于语料库的研究方法有误。乔姆斯基区分了语言能力(language competence)和语言使用(language performance)这两个概念。认为，语言研究的主要目标是建立一种能够反映说话人心理现实的语言认知模式，也就是语言能力模

式。因为只有语言能力才能对说话人的语言知识作出解释和描述。语言使用只是语言能力的外在证据，往往会因超语言因素的影响而发生变化，因此，后者不能确切反映语言能力。乔姆斯基还认为，语料从本质上只是外在化话语的汇集。基于语料的研究所建立的经验模式，充其量只能对语言能力作出部分解释。因而语料不应当是语言学家从事语言研究的得力工具。

2．语料的不充分性。乔姆斯基在《句法理论》一书中首次发现英语短语结构规则具有递归性。这种递归性表明：自然语言句子的数量是无限的，是任何有限的语料所不可能穷尽的。换言之，语料永远是不完整的，不充分的。

转换生成语法学派的上述批评从根本上改变了50年代结构主义语言学的研究方向。在此后的近20年里，整个语言学界几乎唯直觉是从，唯思辨独尊，语料库方法几乎名誉扫地。但是即使在这种情况下，语料库的研究从未完全终止。许多语言学家凭着非凡的学术勇气，顶着无形的压力，继续不懈地从事语料库语言学研究，并不断取得进展。1959年，夸克着手建立“英语用法调查”语料库(Survey of English Usage)。与此同时，Francis 和 Kucera 开始了建造在现在非常著名的布朗语料库的工作，这项工作从开始到最终语料库的建成，花费了近20年的时间。此外，在1975年，Jan Svartvik 在前两项的工作基础之上，开始建造伦敦—隆德语料库(London-Lund Corpus)，并且最终实现了计算机上的 SEU 语料

库。

对此,里奇(Leech,1991)认为:“作为英语口语研究的语料资源,它至今仍无以伦比”。另外,以弗朗西斯(N. Francis)和 Kucera 为首的一批语言学家和计算机专家汇集在美国的布朗大学合力攻关,于 1961 年建立了当今最早的机读语料库—布朗语料库(Brown Corpus)。布朗语料库以共时原则采集不同主题的英语样本,总规模为一百万词次,目的是研究美国英语。这两个语料库可以说是现代语料库的开山鼻祖,并共同为 80 年代语料库语言学的复苏奠定了基础。

三、语料库语言学的复苏时期

80 年代以来,语料库语言学在相对沉寂了近 20 年后开始复苏,并得到迅速发展。主要表现在下面几个方面。

1. 第二代语料库相继建成

80 年代以来,以柯林斯—伯明翰英语语料库(COBUILD)为代表的一大批语料库相继建成。这些机读语料库尽管规模、设计和研究目的各异,但大多采用了较新的 KDEM(Korowai Data Entry Machine)光电字符识别技术,从而使语料的编码和编辑得以从繁重的人工键盘输入中解脱出来。这个时期建设的语料库,不仅规模上有数以十倍的增长,而且建设速度大大加快了,故称第二代语料库。根据美国加州大学伯可莱分校的语言学家爱德华兹(Edwards)在 1993 年

的不完全统计,80年代以来建成并投入使用的各类语料库达50余个,按语种分布如下:

英语	24	法语	4	意大利语	2	丹麦语	2
德语	7	西班牙语	2	芬兰语	2	瑞典语	2

此外,葡萄牙语、南斯拉夫语等语种也相继建立了自己的语料库。在这些语料库中,规模大且特色鲜明的有:

(1)兰卡斯特–奥斯陆–卑尔根语料库(Lancaster-Oslo-Bergen Corpus, 简称LOB语料库)。70年代,在英国兰卡斯特大学著名语言学家里奇的领导下,以研究英国英语为目的,采用与布朗语料库相同的语料分布和采样原则来设计,1983年建成。语料库由五百个样本组成,每个样本约两千词次,总规模也是一百万词次。由于这些特点,人们通常把LOB语料库和布朗语料库视为姐妹库,以进行英国英语和美国英语的对比研究。

(2)法语语料库(Tremor de la Language Françoise, 简称TLF语料库)。该语料库是法国国家科学研究中心与美国芝加哥大学的合作项目。语料的跨度从7世纪至20世纪,包括书面法语各种文体的两千个样本,总规模达1.5亿词次。有关数据已制成光盘,并可通过UNIX操作系统查阅。

(3)赫尔辛基历史英语语料库(The Helsinki Corpus of Historical English)。这个语料库是以罗西尼(Roseanne)等为

首的一批语言学家在赫尔辛基大学建立的。语料包括从公元850年到1720年这一时期的各类英语语篇，并以每百年分段，总库容量达1600万词次。作为第一个历时的英语语料库，它对于从社会语言学、方言学及语用学角度来研究英语的变迁，具有重要价值。

(4)国际英语语料库(The International Corpus of English，简称ICE)。该库于1988年由伦敦大学英语系承建，旨在为世界范围内英语民族变体的对比研究提供数据。语料分别取自所有英语国家，并采用统一的分类和编码系统。每个国家的语料限定为一百万词次，口语和书面语各占一半。语料采样时间限定在1990年至1993年之内。采样对象为18岁以上接受英语教育成长起来的成年人。

在这个时期内，我国语料库的建设队伍也在逐渐壮大，利用语料库进行语言调查和研究的学者也在不断扩大。如利用大规模语料库对汉字和词语的使用频率进行统计调查。见到的主要成果有《现代汉语常用字表》和《现代汉语频率词典》等。本书以后的章节中还将对各种语料库的建设和使用情况进行详细描述，在此不再赘述。

2. 基于语料库的研究项目增多

大批语料库的建成极大地推动了基于语料库的研究项目。下表的统计数字充分说明了这一点。

1959—1991 年语料库研究项目统计表（Johansson，1991）

起止年限	研究项目数
1959—1965	10
1966—1970	20
1971—1975	30
1976—1980	80
1981—1985	160
1986—1991	320

事实表明，计算机语料库是开展大范围语言研究的极好资源，因为它所提供的语料较之先前的材料更具有真实性，其层次结构更加明晰，因而更有助于对语言的不同层面进行描写，对不同语体开展对比研究，进而实现语言的计量研究。

这期间许多研究项目取得了重要成果，有的深化了原有的研究，有的则拓宽了原有的研究领域。如 Halliday (1991)[10]和 Svartvik(1992)[11]等人的概率语法研究；Dottie (1991)的英国英语和美国英语话语风格研究，以及 Sinclair (1985)等人关于英语搭配的计量研究等。

对 80 年代以来英语语料库语言学复苏的原因，近年来多有评说。概而言之，主要有如下两条：

(1)计算机科学的飞速发展与计算技术的迅速普及和应用，为语料库语言学的复苏提供了强大的物质基础。80 年代以来，语料库的发展进入了一个良性循环：计算机的运行速度和存储能力的成倍增长加快了语料库的建设，提高了语料库

的处理能力和处理层次。同时,大量经过标注的语料又反过来促进了语料库的研究和利用。在此期间,诞生了更为先进的研究方法和语言模型,许多先前需要人工处理的标注和统计工作,现在可以通过计算机软件自动或半自动地完成。在这一循环中,计算机显然是一个举足轻重的环节。

(2)转换生成语言学派对语料库语言学的批判,经过20年的实践已证明,有的是错误的,如指责计算机是伪技术;有的是片面的,如对语料库的全盘否定;有的则是正确的,如乔氏关于自然语言句子的数量无限性的观点。对于乔氏倡导的理性主义方法,人们经过跟从、应用和反思之后,也逐渐发现其不足,如不可验证性等。因此,80年代以来语料库语言学的复兴,在很大程度上反映了语言学界的一种普遍心态,即想要恢复语言研究中人工数据和自然数据的平衡。既然语料库研究方法和基于内省的唯理主义方法各有长短,为什么不能让二者共存或结合,以充分发挥其互补的优势呢?为了达到这种有益的平衡,许多语言学家发出呼吁,如:

"语料研究在语言的理论探索中具有中心位置,对语料的开发途径很多,……并非只有一种"。(Halliday 1991)[10]

"从科学方法的角度,语料库方法是一种更为强有力的研究方法,因为其结果是可以验证的"。(Leech 1993)[12]

即使像C. Fillmore这样曾经对语料库语言学有诸多批评的语言学家,也对语料库语言学作出了公允的表述:

"我不认为有这样的语料库:它能包括我要探究的英语词

汇和语法的全部信息,不论它有多大。……[诚然],每每有机会探查语料库,无论多小,总使我获得一些用其他方法无法得到的事实。我的结论是:两类语言学家相互需要”。(Fillmore 1992)

第三节 语料库语言学的发展方向及前景

对于语料库语言学的发展前景,特别是下一个世纪的发展方向,近年来语料库语言学家多有论及。如 Svartvik (1992)[1] 预测“计算机将运行更快,体积更小,价格更低;语料库将规模更大,质量更好,利用率更高。”McEnery(1996)则认为,语料库语言学的发展将主要受语料库规模、类型、国际关注和计算机发展等四方面力量的左右。基于语料库语言学的研究现状,总观各家之说,语料库语言学未来的发展方向将主要体现在以下几个方面:

1. 基础语料库的发展

20 世纪 90 年代以来,由于对民族语言资源价值认识的深化,特别是在欧洲,许多国家的政府或学术机构从维护、发展和规范本民族语言的角度出发,纷纷投资建立大型语料库。如英国国家语料库由牛津大学出版社牵头,参建单位包括兰卡斯特大学、英国朗曼出版有限公司和英国皇家图书馆等。

再如,日本的教育科学文化省于 1989 年组织了三百多位专家,历时 5 年建成了"日语言语数据库",其全部语料已经制成 22 张光盘,其中言语库 19 张,数据库 3 张。这种建库势头仍将持续下去。鉴于大型语料库语料标注工作的滞后,有人认为今后一段时间还应发展小型专用语料库,例如肖特(Short,1996)为研究言语和思维的表达所建立的语体研究语料库。此外,口语语料库的发展应加大力度,以克服目前书面语语料库和口语语料库发展的失衡,促进口语研究的发展。威尔逊 Wilson(1996)预测在不久的将来会有更多的语料存储媒体问世。

2. 语料标注的发展

语料库标注是对语言进行多维、多层面分析的基础,而此种分析结果的受益者不仅限于原标注者,因而语料库的有效利用在很大程度上有赖于语料库标注的层次和质量。为此语料库标注发展的着力点应主要包括:

(1)把语料标注过程中所体现的语言分析规范尽可能写成文件,如约翰森(Johansson,1982)的词类标注规范和桑普森(Sampson,1987)的句法分析规范等。在汉语中,有汉语的分词规范和词性标注规范。

(2)不同分析规范之间应力求调和,即尽可能使用普遍认同的标记;在规范与规范之间进行转换时应提供必要的信息。

(3)目前对语言各层面的标注发展很不平衡。发展较快的层面有词汇层、句法层、语音层和音位层等,今后应重点加

强语义层和语用层的标注。

3. 语料处理工具的发展

语料库分析有赖于计算机环境的支持,即从语料库中检索数据并对语料进行加工的软件工具。充分利用统计学方法,建立科学有效的语料处理工具可以增加语言学研究人员的工作效率。目前软件工具尽管已有了一定数量,但多数工具都是针对某一个特定的语料库,适用范围有限,缺乏通用性。

第四节 计算机在语料库语言学中的作用

手工分析规模较大的文本存在容易出错、难以穷尽和不易重复等问题。虽然几个世纪以来,这种方法在语言研究的许多方面,特别是在词典编纂学上,作出过重大贡献。但是,20世纪中叶计算机的产生,却给基于文本语料的语言研究带来了根本变化。信息革命使我们彻底改变了基于语料的工作方式。无须使用索引卡片和词条工作单,词典编纂学家和语法学家用计算机存储大量文本,就可以快速无遗漏地将指定的词、短语或语块检索出来,并显示在计算机的屏幕上。甚至更进一步,以多种方式对逐词索引的结果进行排序,如按照与指定条目的搭配或其语法行为来排序。

因此，语料库语言学与计算机有着极其紧密的联系。正是计算机使得基于语料库的语言学研究具有高速、可靠、可重复性，以及处理大量文字信息的能力。使用计算机软件不仅大大减少了语言学家在编纂词典和处理大规模语料时单调乏味的手工劳动，而且避免了产生差错的人为因素。计算机除了具有上述对语言条目进行检索、计数和排序等功能外，还能够准确地计算这些条目在文本中的概率。而这些统计数据又可以形成统计语言模型，作为像汉字输入、语音识别、全文检索等自然语言处理应用系统的基础。计算机可以使语言学家超越其直觉，在丰富多彩的大规模真实文本中探寻语言和语言使用的一般规律。基于语料库的语言研究使用量化研究，有助于增强语言描述及其各种应用间的联系。其中，机器翻译、文本--语音转换(TTS)、内容分析和语言教学等应用是最直接的受益者。

实现计算机文本处理的思想要追溯到20世纪70年代。在《文学与语言计算》(Literary and Linguistic Computing)的先驱杂志 ALLC bulletin 上，Govindankutty(1973)发表了一篇工作报告[13]。这篇报告预告了对德威甸语(Dravidian)的处理由人工转向计算机处理时代的到来。他和他的学生经过6年艰辛劳动，实现了用计算机对一篇含30万词次的文本的数据管理和分析。在今天的语料库语言学家看来，使用穿孔纸带的输入方法显得十分古老，这些艰苦的劳动在今天可以直接在键盘上输入。

在80年代中期，语料库语言学家使用大型计算机完成语料的管理任务。在70年代，用一台共享主机，在一个一百万词次的语料库中检索找出“when”这个词目在语料库中的所有出现，需要花费一个小时或更长的时间。到80年代，这样的工作已经能够在几分钟之内完成。现在，即使一台个人计算机，也配有几千兆字节的硬磁盘和几百兆赫的运算速度，光盘已是十分普遍的输入和存储方式。这些均十分有利于文本的存储和分析。

80年代初期，一些热心的计算机专家曾热情地帮助语料库语言学家解决计算机在进行语料分析时遇到的技术问题。到90年代，随着个人计算机的飞速发展和专门为语料分析而制作的软件包的商业化，现在的语料库语言学家终于可以将精力集中在语言学的问题上，而不必再分心去琢磨如何编制程序和使用计算机了。

第五节　语料库语言学的研究内容

语料库语言学的研究对象是文本，这些文本是在语言描述和论证中的证据来源。其中，对语言条目在语料库中分布的计量描述，已经逐渐成为语言研究的一个必要的组成部分。正像里奇(Leech,1992)指出的那样：语言研究的目的是描述语言的使用，而不是语言的能力。正是对使用中语言的观察

导致了理论的产生，而不是相反。

然而，语料库语言学不同于像转换生成语法那样的语言理论，它不是一种语言理论，甚至更不是语言学的一个新的、独立的分支。它是语言学家在进行语言研究时使用的方法。语言学家在研究语言的本质、成分、结构和功能时，需要有一些语言证据来陈述在语言中什么是可能的。在不同的时代，这些证据或来自语言学家的直觉和内省，或来自调查和归纳，或来自对口语和书面语文本的观察与描述。在基于语料库的研究中，这些事实将直接从文本中得到。在这一点上，语料库语言学不同于以语言学家的内省为立论证据的语言学理论。语料库语言学不仅研究语言中哪些词语、结构和使用是可能出现的，而且还要统计它们出现的概率。像其他语言学一样，语料库语言学研究语言的本质、结构和使用，以及语言的习得、语言变异和语言的改变。研究的重点是词汇和词汇语法，而不是纯句法。

1. 语料库的建设与编纂

语料库是语料库语言学研究方法的资源，因此语料库的设计和编纂是语料库语言学研究的基础。它包括语料库的设计、语料的采集、录入和管理等方面。语料库并非是文本资料的简单堆砌，它必须在某种意义上代表一种语言或其子语言。因此，设计一个语料库时，首先要根据语料库建设的宗旨，精心考虑采样原则和样本分布，以便使新建的语料库尽可能地达到预期的代表性。由于在语言的代表性和语料样本的分布

之间还没有定量的关系可循，目前能够参照的仅是前人和自己的建库经验。从这个意义上来说，采样原则和样本分布依然是语料库研究中的一个尚未有答案的开放课题。当然，语料库规模和存储格式的标准化等问题也是设计中必须慎重思考的问题。只有这样才能使建设的语料库真正为语言研究提供系统全面的观察角度和充足论据。

2．语料库的加工和管理技术

主要指用于语料分析、标注、维护和检索的软件工具。语料库不仅仅是文本的集合，它应该具有良好的存取性能，以便使各种研究人员都能从中检索出自己所需要的信息。因此语料的检索是其中一项重要工作。目前，普遍使用的检索技术是逐词索引(concordance)。

随着计算机功效的不断增长，语料库在规模、多样性和使用方便性等方面都发生了巨大的变化，为加工和访问语料库中所包含的信息已经开发了大量软件。但是，作为研究资源的语料库的价值是不能单独用规模来衡量的，应该通过标注给语料库带来“附加”的价值。这就是对语料库进行的多种标注，通过标注明显地扩展了语料库的语言信息含量，从而对各种语言研究作出更大贡献。对汉语语料库来说，词的切分(即分词)是不同于印欧语言的一项特殊的前处理，其后的词性标注、词义标注、句法标注、句义标注以及话语－篇章标注等加工过程，就同印欧语言的情形相似了。

3．语言研究中语料库的使用

“现代语料的巨大包容性及开发语料的种种手段的出现，构成了深化我们对语言认识和理解的强大力量。”(Halliday, 1991)[10]语料库为语言描述提供了丰富的数据资源。在基于语料库的语言研究中，语言学家利用机储语料库去描写语言的词汇和语法。研究内容不仅仅局限于在语言中可能出现什么，还包含它们出现的概率。而有关词汇和语法分布的研究，进一步促进了在文本类型、语言变化及语言变异的研究。通过从大规模语料中抽取信息的技术，语料库为语义研究提供了丰富的上下文信息。

语言研究中的语料库方法所涉及的研究层面及其相关领域很多，包括口语研究、词汇研究、语法研究、语义学、语用学、话语分析和社会语言学等。

4. 语料库语言学在计算语言学中的应用

基于语料库的语言描述的应用是语料库进化中最具有创新性的一项活动。可以将语料库语言学的研究成果直接应用于自然语言处理、语音识别和机器翻译系统中。在进入 90 年代后，为满足实际应用的需要，基于大规模语料库的统计方法在自然语言处理等领域中逐渐占据了主导地位。

综上所述，在语料库语言学中，一部分学者在研究语料库的设计，一部分学者在研究文本分析和处理的方法，另外一部分学者甚至是一大部分学者在研究基于语料库的语言描述及其应用。

第六节 本书的编排

本书的第一章绪论介绍了语料库语言学的基本概念，简单回顾了语料库语言学的发展历史，并且对它的研究内容进行了概述。本书的以后章节将语料库研究的内容逐项展开讨论。

第二章阐述语料库建设。说明在语料库建设中的几个需要慎重考虑的问题，给出建立语料库的方法，针对各种不同类型的语料库来说明语料采集的一些原则，最后介绍目前国内外的一些有影响的语料库。

第三章介绍语料库的加工和管理技术。主要包括三方面的内容：语料库的索引技术——逐词索引；语料库中使用的统计方法；语料库的标注等。其中第三个内容为该章的重点。

第四章讨论基于语料库方法的语言研究。对语言的不同层面采用定量的方法进行描述，以此反映语言在使用中的特性。最后给出一些在汉语和英语中应用语料库方法进行语言研究的实例。

第五章重点介绍语料库语言学在计算语言学中的应用，即将语料库语言学的研究成果应用于自然语言处理的各个领域。

第2章 语料库的设计与开发

从事语料库语言学研究的人员首先面临的任务就是建立语料库。他们必须对语料库应该包含哪些语料以及如何组织这些语料等问题作出决定,并且能够控制以后在使用语料库的过程中将要发生的事情。语言学家则要能够处理语料中的任何语言实例。

第一节　语料库设计和编纂中的问题

语言学家建设语料库的本意是用来调查和分析语言的结构和用法。但是 Johansson (1994)[14]在 90 年代中期已经注

意到与"语料库"(corpus)搭配的最高频动词是"编纂"(compile)。从60年代到90年代,语料库的编纂、结构和规模一直是语料库语言学家们持续关注的主题。

语料库设计与编纂的出发点是:如何使得在其基础上开展的语言调查是合理的和可靠的。因此,Kennedy(1998)[15]指出了语料库设计师所面临的最基本的问题。这就是:这个语料库所采集的语言数据是否能真正代表了某种期望的语言或语体。在语料库的建设和编纂过程中应当考虑的问题包括:一个语料库应当是一种语言的静态样本还是动态样本?它在多大的程度上可以成为一种语言或语体的代表?为了使语料库能满足一般的或特定的研究目的,一个语料库的总规模应当有多大?它应当包含多少个样本以及每个样本应当有多大?下面是他对以上问题的一些讨论。

一、静态与动态

一个语料库可以是以某种方式采集的文本的静态集合,其目的是成为整个语言或在某一特定时期语言的一个代表。第一代一百万词次的语料库就属于这样的语料库。例如,SEU语料库试图以静态方式在不同使用领域的口语和书面语材料中选择英国英语的样本,使语料库可以作为英语共时的代表。在设计这样的语料库时,通常要十分小心地处理如下一些问题,如特定的体裁(writing style),特定的样本规模

等等。1985年夸克(Quirk)[16]等人出版的英语语法大全(A Comprehensive Grammars of English)就是以SEU语料库为基础编撰的。他认为SEU语料库是英国英语的一种快照。这种语料库像是一幅风景照,目的是抓住风景的主要特征。虽然语料库采用了样本式的设计,但是并非所有的语言现象都可以收入这样一个静态语料库。实际上,语料库中只收集了某些主要的体裁。因此,这样的语料库是语言的一系列静态片断。一个样本语料库犹如把一种语言在某个瞬间突然冻结了。但由于设计者采用固定数目的样本和文本类型来小心地加以构造,样本语料库可以方便地同其他构造相似的语料库进行对比。小规模或大规模的语料库都可以是静态的。即使拥有一亿词次的极大规模的英国国家语料库(BNC)也属于这种类型。

对语料库建设的另一种主张是动态的,或监督的语料库(monitor corpus)[17]。一个监督语料库更像是一部动画,而不是一幅快照。这样的叫法是因为,它提供了一种方法来观察语言用法模式随时间变异的情况。大量收集某一时期内的文本,然后通过软件在这些文本中找出与描写目的有关的统计信息,进而对所观察的语言现象作出总结。比如,注意到某些新的结构或词型的出现,或者某些老词型的用法或搭配发生了改变等等。这样一种动态的文本集,将随着新文本的加入而不断地增容和变化。辛克莱(Sinclair)描写监督语料库的这种概念时说,就像随时都掌握着语言的状态。但是,以亿

计的词次数对于任何现实的处理手段来说都显得太大了。由于监督语料库的组成和规模在不断地变化，这使得它不适宜于在不同语料库之间进行对比研究，如比较在不同语言变体中词目的相对频率。所以，监督语料库是一种与静态样本语料库完全不同的语料库。对于一个监督语料库来说，数据的收集通常是随遇的，并且不一定“平衡”。对文本数量的关注取代了对采样计划的精心设计。为了文本的收集、存储和处理，在计算机硬件方面需要昂贵的资源，为了分析又需要复杂的软件和技术经验。因此，对于个别研究人员和众多学术团体而言，用监督语料库来从事研究的机会很小。这种建库方式似乎主要由大型企业、政府部门或专门研究机构所采用。个别研究人员也许要通过付费的方式去访问这种监督语料库。基于语料库的研究犹如大多数语言学研究那样，通常独立进行，很少合作，只是在计算机辅助下通过文本和语言学家的思维的相互作用来完成。因此，如果监督语料库只被少数研究人员所利用，这将是一种很大的浪费。使用监督语料库确实可以为语言学家提供详实的洞察（尤其是对于词典编者和历时语言学家），如关于语言的变迁过程和低频词的用法等。但是，无论如何静态语料库似乎可以保证高频词到中频词的研究，以及音位、词法、句法和许多话语的研究。

目前，随着计算机存储容量和计算性能的不断提高，已经使得用计算机处理上亿次词次的大规模语料库成为可能。并且在许多像语音识别和音字转换等实际问题的应用中，对建

立这样大规模的语料库也有需求。因此,我们认为,当今在建立语料库时,应该尽可能收集所能收集得到的各类语料,建立大规模的语料库。问题是在建立这样的语料库中,应将这些语料分门别类的组织好。使得在具体应用中,能够根据实际要解决的问题,从这个语料库中选择所需要的语料,建立起自己的语料库。比如,若想考察有关领域中词目的使用情况,这时,可以从这个较大规模的语料库中取出有关领域的语料,然后再进行相应的研究。

二、代表性和平衡

同语料库究竟应当是静态的还是动态的问题相联系的另一个问题是:通过选择什么样的文本进入语料库才能达到合理性和可靠性的要求?语料库的代表性和平衡性是一个迄今还没有公认答案的复杂问题。里奇(Leech,1991)[18]曾指出,一个语料库具有代表性,是指在该语料库上获得的分析结果可以概括成为这种语言整体或其指定部分的特性。早期布朗或LOB语料库的结构是经过小心设计的,因此它们通常被分别视为美国英语和英国英语在那一特定时期的代表。当然,代表性和平衡的概念在最终的分析中取决于判断,而且只能是近似的。

其实问题的本质在于:语料库究竟是"什么的代表"?通过几十年来在话语分析和社会语言学的研究,尽管一个样本不足以代表一种特定的体裁或主题,然而由大量各类样本组

成的一个语料库可以成为一种语言的代表，仍然被认为是一个合理的目标。说到底，概括总是科学的一个本质部分。比如在语音学中，尽管用同一种语言说话的每个人的发声都不相同，我们仍然毫无困难地接受有关这种语言的一个语音系统。大型的词典和语法也都是在这种意义上对一种语言作出的概括。

语料库设计中的另一个问题是：在一个一般目的的语料库中，如何达到不同部分之间的平衡？大多数早期语料库都偏爱书面语，对它们赋予很高的权值，甚至只采集书面语的文本。这是因为在现代技术的支持下，书面语的采集仍比口语材料的收集容易得多。即使在规模最大的结构化的第二代样本语料库BNC中，一亿词次中只有10%来自口语资源。与此相反，在规模较小的ICE语料库中，60%是口语文本，40%为书面语文本。这是偏重于口语文本的少数几个语料库中的一个。即使在一个书面语语料库中，应该包括哪些体裁的问题，也不是轻易就能回答的。例如，目前尚不存在一种得到广泛承认的体裁分类方法，更不用说为了平衡究竟应当如何来掌握各类体裁的比例了。

即使被设计的语料库不打算代表整个语言，而只代表一个特定领域、体裁、话题或主题域，也仍然存在一个平衡问题。只有当语料库由一个历史时期中出版的每一件作品组成，由一个作家的全部作品组成，或由任何其他的文本总体组成，平衡问题才能完全回避。在一个语料库中，平衡不能简单解释

为文本的不同来源，比如让口语与书面语的文本总数相等。因为其实没有人知道，任何一天中在一种语言中产生的口语词或书面语词的比例是多少。从个体来说，我们每天接受或产生的口语词大概比书面语词多得多。但是，一篇书面语文本(比如一篇报纸上的文章)也许有一千万人阅读过，而关于买一双鞋的谈话也许除了两个参与者之外不会再有任何人听到了。同样地，在无线电或电视上的一段广播对话比起一位顾客和售货员之间的对话会到达更多人的耳边。

在书面语语料库中，平衡也同样是不容易处理的。辛克莱(1991)建议对一个一般的书面语语料库来说，在选择文本方面的最低准则至少应区别小说和非小说；书本、期刊或报纸；正式的或非正式的出版物；以及对作者年龄、性别和原籍的区分等等。怎样在少数作者与言者之间取得平衡？谁是更权威的？谁产生的文本拥有更多的读者或听众？语料库的设计者想出越来越复杂的方式来力图满足代表性和平衡性的要求。BNC 语料库就提供了采样的经验。

萨默斯(Summers 1991)[19]曾经讨论过为了使一个大型语料库具有代表性而需要考虑的重要问题。她注意到，同从中采样的书面语总体相比，即使一个三千万词次的书面语语料库仍然是相当小的。语料库编纂的经验令她相信，报刊文章的内容和语言同想象类和学术类文章相比，在词汇上是不同的。基于这样的观察，萨默斯主张采用一种"广泛的客观定义的文本类型"来进行初始采样。然后，文本类型的平衡可以

通过对库存语料进行分析来加以修改或微调。她总结了许多选择书面语文本的方法，包括：一种基于学术价值或“影响力”的“钦定”方法；随机采样；作品流通度或文本被阅读的广泛程度，从而偏爱报刊文章和最畅销作品；对于文本典型性的主观判断；在档案馆中文本的可访问性；人们阅读习惯的统计采样；以及依据语言说明来进行文本选择的经验等。当然实际使用的方法是上述这些方法的某种组合，例如从丰富的资源和文本类型中进行选择，用流通度和影响力等因素来指导等等。

另外，从指定年度或期限来选择文本也很重要。历史上的经典文本也许不再被广泛阅读，或不再具有影响力。另一方面，宗教著作，如 King James 版本的圣经(Bible)是几百年前翻译的，但至今仍有影响。

目前，在汉语语料库的建设中，解决语料平衡问题时大部分建设者采用的还是按题材和体裁等来进行的。北京语言文化大学[20]最近在建立语料库时提出了一种依据在应用语言学中流通度的概念建立大规模语料库的方法，从媒体的流通度计算到词的流通度计算，以此为依据建立语料库，以保证语料的代表性和平衡性。

三、规模

怎样使一个语料库具有代表性或平衡，并能够用于对比

目的？问题的关键是语料的质量，但是也有些方面涉及到文本的数量。这些问题不仅关系到语料库的词次总数和不同词型(types)的总数，而且还关系到语料库应该包括多少种文本范畴，每个范畴应该包含多少样本，每个样本应当有多大(词次数，或字次数)等等。诚然，规模和代表性决定了语料库的合法性和可靠性。但是必须再次强调指出，语料库不管有多大，同这种语言的总体相比仍是微不足道的。从这个意义上说，语料库的规模也不是越大越好。

1.语料库的规模

在20世纪70年代，一百万词次的第一代语料库就似乎很大了，在大型计算机上用逐词索引工具进行搜索要化几个小时。到了80年代中期，使用预索引软件，在个人计算机上对一个一百万词次的语料库进行一次逐词索引搜索只需花几秒钟。新一代的语料库，像COBUILD语料库和Longman/Lancaster语料库由于采用了光学扫描等技术，使“录入”和编辑文本比以前容易得多，因此这些语料库的规模也随之增大了。辛克莱(1991)建议，1000－2000万词次可以构造一个有用的、小型的一般语料库，但若要对语言整体作出可靠的描述，这样的规模仍嫌太小。有人甚至认为，规模的受限似乎是语料库的一个固有的缺陷。比如辛克莱(1991)曾指出，即使造出十亿词次的语料库，对于一个大型词表中的大多数词型来说，仍然会显示出相当严重的稀疏信息。其实，这就是著名的齐夫律(Zipf's Law)。这条定律称，若按照词频 f 由高到低

的排列顺序给语料库中每个词(型)指派一个由小到大的整数秩 r(rank),r = 1,2,3,...,则 f 与 r 近似成反比,即

$$f * r = k$$

或

$$f = k / r$$

其中 k 为常数。不难看出,词(型)的频率 f 与其秩 r 之间保持一种双曲线的函数关系。对不同语言的许多语料库的调查都证明了齐夫律的普遍性:极少数高频词(型)的出现次数已覆盖了一个语料库总词次数的绝大部分,而词(型)总数中大约一半的词(型)在这个语料库中却只出现一次。近年进一步的调查表明,齐夫律不仅适用于一种语言的词汇分布,而且也反映了句法规则的分布状态。一方面,极少数最常用的句法规则就已经覆盖了一个语料库中绝大多数的句法结构现象;另一方面,很多规则在这个语料库中只出现过一次。有趣的是随着语料库规模的扩大,句法规则的数目呈不断增长的态势。语料库的这一统计结果向乔姆斯基的著名假设——一种语言(句法)规则数目的有限性和句子数目的无限性,提出了挑战。

对一个语言项来说,为了达到描写的充分性,究竟需要多少个标记(tokens)呢?统计显示,在一个典型文本(或一个百万词次的语料库)中,大约有 40% - 50% 的词型只出现一次。朗德尔和斯托克(Runddle and Stock 1992)[21]在他们为基于语料库的词典学所作的综述中注意到,在 Longman/Lancaster 语料库中,词"break"尽管出现了 8267 次,但像"break"用作

"news breaking"(消息传开)等用法的出现次数仍嫌不足,不能为词典编者提供必要的信息来判断它是否适合作为词典条目。但是,如果出现一次太少,那么为了发现搭配、多义词和词法特性等语言现象,一个指定词型或词义的出现频率要多高才够用呢?例如,通过LOB语料库和布朗语料库的调查,词"circumstance"在美国英语和英国英语中约90%的出现都是复数形式(circumstances)。另一方面,英语中像"at"这样的高频词,在一个一百万词次的语料库中统计到5500个标记。对于大多数描写目的而言,这个数字已足足有余了。对于词典学或词汇语法研究来说,如果数据太多,所需的人工分析将难于应付。一个词(型)如果在逐词索引中有超过1000个标记出现,这对词典编者来说已经是数据分析的最高极限了。为此语言学家和词典编者要求有一种软件来支持他们,以便从大量检索结果中获得最佳采样,并且对如何进行分析提供有效的技术指导。这是在大型语料库中分析高频词时出现的情况。由此说明,语料库规模太大,也会对语言分析带来负面影响。一般来说,这个问题可以通过对检索结果的随机采样来解决。但是,对于在一个语料库中只出现一次的那些词(型)来说,需要为这些词(型)专门建立一个子集。然后,再到规模更大的语料库中去搜集这些罕见词(型)的用法实例。

调查显示,在一百万词次的LOB语料库或Wellington语料库中,大约有100个词(型)出现次数超过1000次。而在一个规模为一亿词次的语料库(BNC)中,出现次数超过1000次

的词(型)也相应增长到约8000个。这8000个词(型)大约覆盖了语料库中95%的词次。其余的5%词次(大约500万词次)可能由50万或更多的词(型)组成。因此,为了研究一种语言的词汇现象,尤其是要想对低频词现象作出充分描写(如搭配),极大规模语料库是必要的。但是如果无法同大量检索出来的数据打交道,拥有越来越大的语料库也并非好事。即使一个语料库拥有海量的文本收集,如果设计不善,也不一定就可以通过它对一种语言作出概括。在这个意义上来说,一个巨型语料库不一定能比一个较小的语料库更好地代表一种语言或它的一个变体。在目前阶段,还不能确定对于一般目的或特定目的来说,一个语料库究竟要多大。因此,与其过分关注一个语料库的数据规模问题,编者和分析员不如对数据的质量给予至少同样的关注。

文献[15]认为:就英语而言,对于节律(prosody)研究来说,为了对大多数描写项作出概括通常十万词次的语料库已经足够大了。为了对动词用法进行可靠的分析,可以在一个五十万词次的语料库上完成。许多句法结构和高频词汇的研究一般要求语料库的规模为五十到一百万词次之间。像BNC这样结构化的一亿词次的语料库,可以通过与不同结构的较小语料库,如布朗、LOB语料库或ICE语料库的分析结果进行对比,来解决与规模和代表性有关的问题。这样做可以找出规模、文本范畴或总体结构等因素对特定语言项或层面使用情况的影响程度。一般来说,在任何文本范畴内,通常样本数

越多,分析结果的可靠性也就越高。

2. 样本的规模

与语料库规模有关的另一个问题是样本的大小问题。在布朗和LOB这样的第一代语料库中,总共有按照指定体裁的百分比随机选取的500个样本,每个样本不少于2000词次。其中有些样本是整个文档,而大多数样本只是原有文档的一部分。由于在绪论和结论等章节中摘录的文本具有不同的话语特性,因此对只截取文档一部分的采样方式来说,如果采样方案不周全,可能会破坏语言的整体面貌。反之,在结构化语料库中若采用完整文档,将导致样本数的减少,从而会影响语言的多样性。只有当语料库规模足够大时,这样的问题才可以避免。

比伯(Biber 1993)[22]研究了LOB语料库和LLC语料库中55“对”样本之间10个语言特征的分布。这些样本分别来自范围广泛的口语和书面语体裁。他从每个样本中截取属于同一文档的1000词次。然后比较每对样本之间的内部变异。他的结论是:为了代表样本的文档范畴,样本规模在2000-5000词次左右已经足够大了。进而,他认为像LOB那样的语料库,每种体裁的样本数(20-80个文本不等),对于通常进行的各类变异研究,即基于相关性的分析来说,是合适的。

第二节　建设一个语料库

文献[23]指出：对一个语言学家来说，给出一个语料库的说明，阐明语料库中要采集的文本范畴或类型，以及每一文本范畴在整个语料库中所占的比例，是一项很复杂的工作。这件事情最好应该由社会语言学家来完成。而语言学家的任务应当是分析和描述语料库中出现的语言实例。但目前实际情况是，语言学家或计算机专家也在做着这种语料库样本分布的设计工作，这是有点强人所难了。

在筹建语料库之初，首先要考虑的是建立语料库的目的：是一般用途的语料库，还是特殊用途的语料库。一般语料库应为各种语言研究提供大量好的语言实例；特殊语料库可能是为某种自然语言产品服务的，具有明显的领域针对性，因此是该产品开发过程中的一个阶段。本章只限于介绍一般语料库的构造。

一、语料的来源

使用已有的输入技术，可以采用多种方式来收集语料——制作电子文本，或利用已有的电子文本。

1．制作电子文本

有两种方式可以帮助制作电子文本，它们分别是：

(1)光电扫描输入。目前,在光电扫描输入中印刷体识别技术已逐渐成熟,因此,大量的印刷书籍的输入可以采用这种输入方法。这种方法可以免去再用键盘方式对书籍录入。例如:现在市场上出售的《四库全书》光盘版,就是采用这种方法输入计算机的,他们在光电扫描输入后,采用了清华大学计算机科学与技术系的印刷体汉字识别系统进行识别[24]。但是在使用光电扫描的输入方法时,首先需要保证这种扫描系统具有较小的识别错误率,其次就是要对输入的文本进行人工编辑,以改正扫描中的错误和制作文本标签。

(2)键盘输入。通过键盘将文本手工输入计算机。对光电扫描输入难以识别的文稿,应采用键盘输入方式。这类文稿包括传单、手迹和语音记录等等。有时,如果光电扫描器不是十分有效,与其花费大量的人力去纠正扫描中的错误,还不如请一位熟练的录入员用键盘整个地输入这样的文本。汉语最初在语料库建设时使用的就是这种方法。比如在80年代初期,山西大学计算机科学系受国家语委的委托,进行的常用汉字的字频统计工作就是在这种输入方式下完成的[25]。由人工输入一个月的《人民日报》,然后计算机来自动完成字频的统计工作。

2. 使用现存的电子文本

现在,像报纸、杂志和书籍等出版物已经普遍以电子文本的形式存在。利用这些文本是建立语料库的一种最直接的方

法。此时,语料库建设者的工作就是对这些文本进行格式转换,将它们转换成语料库指定的文本格式。

大多数语料库建设项目都需要同时使用以上两种文字输入方式,因为其中的每种方式分别适合于某种材料的输入。例如,手写的文档和语音的转写文本适于用键盘输入,而大多数报刊、杂志本身已经以电子文本形式存在。对于那些采用传统方法出版的书籍,使用识别正确率高的光电输入系统,不失为一种快捷的输入方式。

二、语料库的设计

语料库的设计包括决定语料库的规模、库藏语料的文本范畴、各文本范畴在整个语料库中所占的比例、建立可查询的分类体系,以及文本的实际采集等。文献[23]对语料库的设计问题进行了较详尽的描述,下面简单介绍语料库设计中需要慎重确定的几件事情,以及作者的一些观点。

1. 语料的分布

从本质上来说,语料库就是由各种类型的文本组成的。因此,选择何种类型的文本对语料库的建设具有重要的影响。下面是几个首先要考虑的问题。

(1)口语和书面语

在建立语料库时,首先要确定它是一个书面语语料库,还是一个口语语料库,或者是两者并存的语料库。许多语料库

都只含书面语文本，这将使语料库不能全面反映出这种语言使用的真实状态。许多语言学家都认为，对语言构造的基础研究来说，口语比书面语更具指导意义。没有任何书面语文本可以完全取代口语。

但是，比起书面语文本，口语语料的采集要困难得多。虽然像电影脚本、剧本、会议记录、法庭案例和电视广播这样的语言材料是不难采集到的，但它们是经过“修饰”后的语言，不可避免地带有一定的人工痕迹，不能完全反映出自然会话中的语言精髓。人们通常把这种口语称为准口语（Quasi-speech）。

汉语的语料库建设中，收集的多是书面语文本语料，如报刊、书籍等，在口语语料的采集方面的工作做的较少，尤其是人们的自然会话。

(2)正式的和文学的语言

语言材料可以选自多种语言形式，既可以是正式的或非正式的语言，也可以是文学的或普通的语言。正式的语言比非正式的语言容易获得，文学语言比普通语言更容易被人们注意。那些短暂的、非正式的书面信函或日记，虽然常被人们所忽略，但它们才真正地代表了被广泛使用的普通语言。同样，在语料库中没有必要收入所有的当代文学作品，只收其中少数一部分作品就可以了。

(3)语言的典型性

语料库最基本的用途就是去发现隐含在语言中的最本

质、最典型的东西。如果在一个语料库收藏的语料中,一大部分被一位作家的文学作品所占据,那么这个语料库在很大程度上反映的只是这位作家的写作风格,它在一般意义的语言研究上就失去了价值。

同样,在报刊杂志中,不同的作者具有不同的写作特点,在语料库中尽可能收集多种文本类型的报道对语言研究非常有益。如果想让语料库更接近于现实的语言,那么语料库就必须尽可能地包含大量普通作者的作品。

(4)文本的产生时期

大部分语料库都试图涵盖某一特定时期的作品,使其能最真实地反映这一时期的语言使用状况。这样的语料库叫做共时语料库。与此相反,为了对一种语言在不同历史时期在词汇和语言结构方面所经历的变异作出系统的研究,就需要采集不同时期的作品来组成所谓的历时语料库。

2. 语料的规模

在确定语料的规模时,需要着重考虑如下几个方面的因素。

(1)语料库的大小

语料库的大小是研究者在建立语料库时首先关心的问题。在语料库研究的早期阶段,由于电子形式的文本很难收集,只能靠人工录入有限规模的文本。因此语料库的规模比较小。在这种情况下,为了反映语言的普遍特性,就需要小心地设计语料的分布。从长远观点来看,语料库的规模将会随

着计算技术的进步而不断增大。这是因为齐夫律告诉我们,在一个语料库中词频的分布通常是极度扭曲的。比如通过对英语的调查,人们发现出现频率最高的词是"the",它的频率约是紧跟其后的两个词"of"和"and"的词频的两倍;之后,词频迅速下降,到第19个词"be"时,其词频是"the"的词频的不到10%,而第84个词"two"的词频是"the"的词频的5%。

因此,如果用语料库方法来调查一种语言的词汇时,为了能最大程度地覆盖这种语言的词汇,就需要建设尽可能大的语料库。

(2)样本的大小

究竟语料库中的每个样本应该包含多少个词,一直是语料库建设者探讨的一个问题。对这个问题,不同的语料库建设者有不同的观点。例如在LOB语料库中,每个样本的大小都不少于2000词次。虽然这种做法后来被许多设计者仿效,但也有的批评者认为,一个如此规模的样本不足于反映原始文本的语言特性。比如,新闻短讯(如一句话新闻)的写作方法肯定有别于记实性报道的写作手法,而由于新闻短讯的文章大小常常不足2000词,从而不会收集到语料库中。从而从语料库中统计得到的语言特性并不能涵盖新闻短讯的语言特性。再者,对于长篇小说来说,只收集其中的2000词次的样本也不能够反映这篇小说的语言特性。因此为了弥补这方面的缺陷,一般应从选定文本的全文中随机地实行采样,以免采集的样本集中在原始文本的某些段落上。

因此在条件许可的情况下，采集整个文本不失为一种好的收集方法。这种方法不需要担心原始文本中各个段落之间的语言差异。语料库中的每个样本都是一个完整的文本，这比只收录文本的一部分更好，因为它可以提供更广泛的语言研究，而且不必担心采样方案的合理性。关键是在建设这样的语料库时要能够设计出好的语料库管理系统，使得语言学家在使用语料库时做到"各取所需"。

三、设计存储系统和保存记录

语料库建设的目的就是要利用运用中的语言事实对语言进行研究，因此小心地设计语料库的存储格式，标注语料中的各种有用信息是语料库建设者的又一项重要工作。对英语而言，存储一个一百万词次的语料库需要 8 到 10 兆字节的磁盘空间。如果对这个语料库进行语法标注，则另外需要 3－5 兆字节的空间。如果对这样规模的语料库进行句法分析标注，大约还需要 30 兆字节的磁盘空间。随着计算技术的发展，计算机的存储不再是语料库建设的一个问题，如一张光盘就可以存储百兆字节的信息。但是只有当语料库中的文本以及与这些文本有关的信息能够方便地存取，这个语料库才会有意义。因此，在编纂一个语料库时，首先应该设计一种系统的索引方法，将存放在计算机中的文件和原始的声音和视频信息联系起来。除此之外，还应该保存文件的目录以及所有的文

件,并且要将它们与工作的拷贝分开存放。

另外,由于在建立语料库时需要采用不同来源的语料信息,如果采用不同的方法对这些文本进行编码和设定标志信息,将导致计算机语料库管理上的混乱。因此为了统一管理这些语料,需要使用一些文本中的特性作为标记,常用的表示文本特征的信息包括行分割符、行数、章、节、段落标志等。在一个语料库中,如果这些标志设计得不一致,计算机就不能正确识别这些标志文本结构的有用信息,从而无法区分真正的文本和编码,产生错误的输出结果。

在20世纪80年代,为了增强电子文本的可移植性,避免不必要的重复输入工作,出版业研制了一种对文本实行电子编码的国际标准。由此,具有灵活性和独立性的软件系统——标准的通用标记语言(The Standard Generalized Markup Language,简称 SGML)已经被越来越多的同行采用来进行电子文本的编码。

还需要说明的是,在将收集到的各种形式的文本变成语料库统一的文本格式之前,应该首先取得这些文本拥有者的许可,以免引起不必要的版权纠纷。

四、语料库的维护

语料库一旦建立起来以后,其中总有许多错误需要修正,或者要对语料库进行改善,因此需要对语料库进行日常的维

护和升级。这样才能适应新的软硬件和用户需求的改变。另外，有关语料库的检索系统、语料库的处理和分析工具，也越来越引起人们的注意。可见，语料库的维护已变得越来越重要。

第三节 语料库的类型

存在着不同类型的语料库是大家都承认的事实，但是用什么样的术语来描述它们，却并无公论。沃克(Donald Walker)[26]曾用以下四个术语来区分语料库的类型，但在这个问题上还有许多有待澄清的争议。

1. 异质的(heterogeneous)

这是一种最简单的语料收集方法，尽可能广泛地接受各种材料，没有任何事先规定的选材原则。ACL/DCI 语料库就是这样的例子。另一个例子是牛津文本档案库 OTA，它收藏的文本在格式和内容上各异，它们的存储格式和原来的出版物完全一样。

2. 同质的(homogeneous)

它是“异质”的对立面。美国政府的 TIPSTER 项目提供了这样的一个实例，这个语料库只收军事方面的文本，如设备故障、危险性评估及其他类似科目。为个别作家建立的作者语料库也属这种情况。

3. 系统的(systematic)

系统性采集是为了保证语料具有广泛的代表性。布朗、LOB、AHI和英国国家语料库BNC都是遵循这一原则来收集语料的。这种语料库在建立时,充分考虑了语料的动态和静态问题、代表性和平衡问题以及语料库的规模问题。

4. 专用的(specialized)

存在各式各样的专用语料库,如北美的人文科学语料库,卡耐基梅隆大学的CHILDES语料库。

第四节 国外语料库介绍

语料库语言学的宗旨就是:通过大规模真实语料的调查,来发现并总结自然语言的各种语言事实和语言规律。国外从20世纪60年代就开始了语料库建设。30多年来,世界各国政府、企业集团和学术团体已经建成或正在建设着各式各样的语料库。本节将介绍国外几个影响较大的语料库。

一、SEU语料库

1959年伦敦大学夸克(Randolph Quirk)组织发起了"英语用法调查"[27](The Survey of English Usage,简称SEU)项目,有计划地收集不同语体的大量语料,并利用计算机对收

集到的语料进行储存、分类。语言科学史上的第一个大型计算机语料库从此产生。SEU 收集的语料包括书面语材料和口语材料。在这两种语料中都收入了不同范畴(或类型)的文本。该库的各种语料成分及其分类见表2-1。

表2-1 “英语用法调查”语料库的结构

原始书写语料(100篇)					
印刷品(46)		非印刷品(36)		口语(18)	
人文科学 6		连续书写品	想象类 6	剧本 4	
自然科学 7			资讯类 6	正式演说 3	
教学 6		社交书信	亲密 4	广播新闻 3	
报刊	一般新闻 4		平等 4	谈话	资讯类 4
	专业报导 4		疏远 4		想象类 2
文书 4		非社交书信	平等 4	故事 2	
法律 3			疏远 4		
论说文 5		日记 4			
散文小说 7					
原始口说材料(100篇)					
有准备的演说		6		不公开	亲密 24
自发言论	演说	10			疏远 10
	评论	体育 4	交谈	可公开	亲密 20

（续表）

		其他 4			疏远 6
				电话	亲密 10
					疏远 6

可以看出，该库共收集200个语篇，口语和书面语各占一半。每个语篇500字左右。整个语料库的库容量为一百万词次。内容包括了各种不同语体和社会的各个层面。

纵观国外语料库的发展历史，夸克的SEU语料库，无论是在研究观念上还是在方法上，都是一大创新。它为语料库语言学开了个好头，为语言学研究提供了全新的科学手段。

二、布朗语料库

20世纪60年代，弗朗西斯（Francis）和库塞拉（Kucera）在美国布朗（Brown）大学建立了世界上第一个根据系统性原则采集样本的标准语料库——布朗语料库。建库的主要目的是研究当代美国英语。布朗语料库规模为100万词次。这是一个按共时原则采集文本的语料库，只选录1961年间由美国人撰写出版的普通语体的文本。全部语料分成15种体裁，共500个样本。每个样本不少于2000词次。根据布朗语料库的统计数据，1967年布朗大学出版社出版了当代英语词频词典[28]。70年代，格林（Greene）和鲁宾（Rubin）设计了一个TAGGIT系统，用来对布朗语料库的百万语料进行词性标

注，他们设计的词类标记共 81 种，上下文约束规则总计 3300 条，自动标注的正确率达 77%[29]。

1．布朗语料库的语料分布[30]

布朗语料库中的语料分 A—R 共 18 种类型，其中 A—J 类属于资讯类语体，而 K--R 类属于想象类语体，其中数字表示各类体裁中样本的数目。

A．报刊：新闻报道

	日报	周刊	总数
政治	10	4	14
体育	5	2	7
社会	3	0	3
现场消息	7	2	9
商业	3	1	4
文化	5	2	7

B．报刊：社论

	日	周刊	总数
社会事业性的	7	3	10
个人的	7	3	10
给编辑的信	5	2	7

C．报刊：评论　14　3　17

（评论的对象是戏剧、书籍、音乐和舞蹈）

D．宗教

书籍	7

期刊	6
传单	4
E. 技能和爱好	
书籍	2
期刊	34
F. 流行的传说	
书籍	23
期刊	25
G. 纯文学、传记、自传等	
书籍	38
期刊	37
H. 混杂的	
政府文档	24
基金报道	2
工业报道	2
大学目录	1
工业机构	1
J. 教科书	
自然科学	12
医学	5
数学	4
社会和行为学	14
政治、法律和教育	18

人类学　18

工程技术　12

K．一般小说

小说　20

短故事　9

L．侦探小说

小说　20

短故事　4

M．科幻小说

小说　3

短故事　3

N．冒险或西部小说

小说　15

短故事　14

P．浪漫爱情故事

小说　14

短故事　15

R．幽默

小说　3

随笔　3

在语料的体裁、子范畴和样本数确定下来之后，样本通过随机采样方法得到。首先从各类体裁目录中按样本数要求随机地选出进入语料库的文本，然后从选中的文本中随机截取

不少于2000词次的片断作为一个样本,采样时还要保证文本的最后一个句子必须是完整的。指派给每个文本的编号,是在上述类型号后面再加一个两位数字。

2. 不同的版本

在布朗语料库建成后,陆续推出过六种版本,分别供各种研究目的学者使用。这些版本分别是:

版本A:是布朗语料库最原始的形式,建成于1964年。受那时计算机打印设备的限制,语料库使用了较复杂的编码技术。

版本B:这是对版本A进行处理后形成的版本,处理中删除了语料库中的标点符号、连字符和公式符号。因此该版本被称为"被剥离的版本"(stripped)。这样做,有利于对单个词的研究,并且这个版本为弗朗西斯和库塞拉进行现代美国英语的频率统计[30]做好了准备。

版本C:一个带有词性标记的版本(tagged)。对版本B中部分文本进行标注,其中每个词都被指派一个唯一的词性标记。在这个标记集中,共有81个语法标记。

卑尔根I:这个版本以及后一个版本都是在卑尔根大学的挪威人文科学计算中心乔斯廷(Jostein)的指导下完成的。文本中包含大小写字母、一般的标点符号标记和最少量的代码,并且保留了排版信息。

卑尔根Ⅱ:这一版本与前一版本的不同之处是减少了排版信息,并有缩微胶片和完整的逐词索引系统。

布朗MARC版本：这一版本是在斯坦福大学完成的。目的是使布朗语料库能够与两个通用的检索软件兼容。这两个软件分别是：从语料库中检索出含有指定词目或部分上下文的完整句子；根据不同关键词及其上下文生成索引行文件。

布朗语料库从语料库的整体规模，语料的分布和语料的采样上都经过了精心的设计，一致被公认为是一个能够反映语言共性的平衡语料库。

三、LOB语料库

70年代初，由英国兰开斯特(Lancaster)大学里奇倡议，由挪威奥斯陆(Oslo)大学约翰森(S. Johansson)主持完成，并最后装备在卑尔根(Bergen)大学挪威人文科学计算中心的LOB语料库[31]，是布朗语料库的姊妹库，目的是研究当代英国英语[32]。为了能同美国英语进行对比研究，它的规模和分布方案都与布朗语料库类似，表2－2给出这两个语料库的语料分布情况。

表2－2　布朗和LOB语料库的结构

编号	语料的体裁	样本数目	
		布朗语料库	LOB语料库
A	报刊：新闻报道	44	44
B	报刊：社论	27	27

（续表）

C	报刊：评论	17	17
D	宗教	17	17
E	技能和爱好	36	38
F	流行的传说	48	44
G	纯文学的、传记和自传	75	77
H	混杂的	30	30
J	教学	80	80
K	一般小说	29	29
L	侦探小说	24	24
M	科幻小说	6	6
N	冒险和西部小说	29	29
P	浪漫爱情故事	29	29
R	幽默	9	9
总数		500	500

兰开斯特大学的研究小组为这个语料库设计了含有133个标记的语法标记集，他们采用与布朗语料库不同的词性标注方法，也对LOB语料库进行了词性标注。他们研制的词性标注系统TAGIT，利用带有词性标记的布朗语料库，通过统计获得一个反映任意两个相邻标记同现的转移概率矩阵，根据这种统计信息而不是规则对一百万词次的LOB语料库进行了词性标注，一下子把标注准确率提高到96－97%。这是语料库方法在自然语言处理领域所取得的一个里程碑式的成

功。许多从事语言信息处理的研究人员正是从这个事实中，清醒地认识到过去基于直觉的人工智能方法的局限性，以及基于观察的统计语言模型的强大生命力。毫不夸大地讲，正是TAGIT系统在语言处理领域开创了崭新的理念和方法，为90年代语料库方法的繁荣吹响了大进军的号角。在带有词性标记的LOB语料库的基础上，1989年约翰森(Johansson)和霍夫兰德(Hofland)编辑出版了英语的词频和词类频率统计结果[33]，随后还将出版英语词类同现频率统计结果。英国兰开斯特大学和利兹(Leeds)大学的研究人员还对LOB语料库进行了句法标注，以便用它来支持基于概率模型的自动句法分析。

四、LLC口语语料库

以上三大语料库的建立，结束了个人费时费力地收集语言材料的历史，确立了语料库语言学在语言研究中无可争议的地位。三个语料库所收集的语料已经从早期的词语、短语、单句扩大到语篇；收集的范围从特定语言扩大到方言和语言的其他分支。但是这三个语料库尤其是后两个语料库收入的语料均为书面语，无法提供有关口语的材料，因此，1975年开始了口语语料库的建设。

60年代伦敦大学著名语言学家夸克(Quirk)组织的英语用法调查课题曾收集过2000小时的对话和广播等口语素材，

并随后整理成书面材料。这些材料后来由瑞典隆德(Lund)大学斯瓦特维克(Svartvik)教授主持全部录入计算机。他们当时发起并组织了一项“英语口语调查”(The Survey of Spoken English,缩写为 SSE)[34]。这项工程实际上是 SEU 的姊妹工程,目的是以计算机自动化处理方式获取 SEU 语料库的英语口语原始资料。语料库的标注包括节律分析、语调单位、重音、语调等,这是很有价值的英语口语研究资源。

SSE 工程于 1981 年完成。这个口语语料库被称为“伦敦—隆德口语语料库”(London-Lund Corpus of Spoken English,缩写为 LLC)。LLC 最初包含 87 个文本,每个文本约 5000 字左右。为了检索方便,首先对它们进行了详细的分类编目。这些文本被分为五大类,包括:

1. 面对面交谈;
2. 电话交谈;
3. 讨论、采访、辩论;
4. 未经准备的当众评论、论证、演讲;
5. 经准备的当众演讲。

接着,又在这些分类后编上子目录。最后用字母 S 和数字给每个文本都加上了标志。斯瓦特维克除了给文本中的每个语段标出语调及节律外,还精心设计了一套索引程序,叫做关键词居中索引(key word in context,简称 KWIC)。这样一来,不仅为索引某个文本提供了方便,还可以用这套程序检索某个段落,甚至某个词在整个文本或段落中所处的位置、搭配关

系、属何种词类、出现次数等。这就要求不仅对每个段落编码，而且还要设计相应的词类标记。在词类标记中，先用不同的英语大写字母来表示不同的词类，然后在每个大写字母后附加其他符号以表示词的不同变化形式。例如，在名词 N 的后面加上 +2 来表示名词的复数形式，用 N+z 表示名词的所有格等等。另外，为了对口语语法进行更为精确的研究，还设计了一套语法标记，以区分句法分析单位。他们还实现了一个短语分析程序。LLC 语料库的最终规模达到了 50 万词次。

五、COBUILD 语料库

COBUILD(Collins Berminhan University International Language Database)是英国科林斯(Collins)出版社和伯明翰大学联合建造一个英语语料库，目的是在语料库的支持下从事词典学研究。80 年代，在辛克莱(John Sinclair)教授的带动下，建立了现代英语语料库 COBUILD，并在此基础上研制开发了语料数据管理和分析的软件工具，形成了一支语料库语言学和词典编纂学的专家队伍[35]。

COBUILD 语料库在 1980 年指定的选材原则是：

1. 书面语占 75%，口语占 25%；
2. 收入的语料应是广为流传的“标准英语”，不收方言。

   其中英国英语占 70%。美国英语占 25%，其他地区

的英语占5%；

3. 反映当代英语的用法，收录的材料尽可能新；

4. 不收诗歌、戏剧和科技方面的语料；

5. 只收年龄在16岁以上的成年人语料，女作者的比例不低于25%；

6. 入选的语料不是样本或片段，而是平均长度为7万英语词次的全文或长篇选录，以利于在篇章层次上进行语言研究。

最初建立的语料库规模为2000万词次。正是在这个大型语料库的支持下，1987年英国科林斯出版社出版了《Collins COBUILD English Language Dictionary》[36]，使全世界词典编辑界耳目一新。COBUILD词典与众不同的地方在于，它在选词、用法和释义等方面拥有最翔实和定量的证据，所有例句来自真实语料而不是编者生造出来的。他们的经验是借助语料库及语料分析方法来进行大规模语言学研究的又一范例。

COBUILD语料库主要用于研究现代英语的词汇、语义、语法和用法。随着一些新的语言材料不断地加入语料库中，使COBUILD语料库成为一个动态的语料库。从对COBUILD语料库过去十年的研究发现，英语的词汇非常庞大并且用法多种多样。因此语言研究必须有足够多的样本。目前这个语料库被称为“The Bank of English”，语料规模为3.2亿词次。这一语料库已经经过词类标注，并且有近2亿

词次的语料已经进行了句法分析。该语料库目前的语言材料都是较新的，大多数是1990年以后产生的文本。书面语材料包括：小说和非小说类的书籍、报纸、杂志、指南、传单和报告等；语音材料包括：日常谈话、广播、会议、接见和讨论等。辛克莱认为这个语料库提供了大多数人在日常生活中听说读写英语的客观实例。

COBUILD还为词典编者和语言学家提供了一个对语料进行复杂分析的软件。它可以完成以下功能：搜索特定词的组合模式、查找一个词的词频、找出一个词的使用实例并进行分析、把检索结果拷贝到硬盘上。在信息技术时代，计算机需要处理越来越多的语言数据，包括文字处理器、拼写检查、机器翻译等，而计算信息服务是其中一个关键部分。COBUILD语料库可以提供有关词汇和语法的丰富信息，来改善上述这些技术。

六、朗文(Longman)语料库

朗文语料库的最终生成是朗文语料库委员会(Longman Corpus Committee)从1988年1月至1990年11月的工作结果。萨默斯(Summers)在文献[37]中指出了朗文语料库在设计上的几个特点。

1. 朗文语料库的目标——遵循客观准则，构造多用途语料库

朗文语料库的目标是根据一份清晰的、适当构思的规范来收集大量文本，建立一个全新的英语语料库，用以编纂词典和供学术界使用。早期的布朗和SEU语料库，都是根据事先确定好的规范来收集文本的。语料库的组成通常（至少部分地）反映了语料是可以直接拿到的材料，而不是设计者原先要选择的理想材料。这种“随遇的”方法会导致产生“扭曲”的实例。朗文语料库的设计方法完全不受材料的可获得性的约束，它的主要结构使它不同于那些早期的语料库。

2．朗文语料库的设计原则

（1）本族语言者的直觉和语料库的权威

朗文语料库的设计者充分信赖本族语言者的长期经验，对本族语言者的直觉给以很高的评价。依靠这种直觉，词典编者能分析和解释语料库的生数据，并且区分什么是典型的，什么是罕见的。但与此同时，语料库也为词典编者或语言学家提供了大量比他们直觉更客观的数据，时常会改变人们过去对某些词语或语法模式的误解，并且揭露出词语的许多新的特性。这是甚至受过最佳训练的词典编者也未必能想到的。在语料库的建设中，重要的是既要相信本族语言者的直觉，也要重视语料库的权威性。

（2）向研究人员提供语料库

目标：设计一个反映20世纪英语的平衡语料库，覆盖美国和英国英语，也包括英语本族语言者的其他主要变体，包括书面语和口语。

主要用途：提供客观的语言数据，根据它可以对词和短语的意思及典型行为作出可靠的概括，并以此作为词典、语法和各类语言著作的基础。

次要用途：提供 20 世纪英语的一个大规模平衡语料库。

(3) 首先开发书面语部分

3. 语料的选择方法

采集的语料必须反映始于 1900 年的 20 世纪英语，当代的材料更多些。文本类型分为知识性(informative)文本和想象性(imaginative)文本，它们所占的比例分别为 60% 和 40%。实际上，小说相对于非小说来说，在整个语言中的所占比例不足 40%，但是朗文语料库的建设者认为，与非小说文本相比，小说是一种更有影响力的体裁，而且有更多的读者。这个事实可以通过查阅图书馆的借阅统计数据得到。

(1) 主题域：话题胜于体裁

朗文语料库选定了 10 个分布广泛的域，叫做超级域。其中既包含占总数 40% 的小说类，也包括另一个分立的范畴——诗、戏剧和幽默。语料库的书面语材料是话题驱动，而不是体裁驱动的。这十个超级域及其在语料库中的比例分别为：

1)自然和纯科学	6.0%
2)应用科学	4.3%
3)社会科学	14.1%
4)世界事务	10.4%
5)商业与金融	4.4%

6)艺术 7.9%

7)信仰和思想 4.7%

8)休闲 5.7%

9)小说 40.0%

10)诗歌、戏剧、幽默 2.3%

总数 99.8%

(2)主要的文档特征

为了把语料库中的文本分类为文档类型,采用了四个外部因子作为主要特征:地域、时间、媒介和等级。其中除了“等级”以外都是客观性度量。书面语中的每个文本都要指派这些特征。

a 地域:涉及英语的主要国家。百分比为:英国 50%,美国 40%,其他国家 10%。

b 时间:朗文语料库没有采用共时语料库的思想,而遵循了历时语料库的语料采集方法,覆盖了1900年以来的英语材料,所以更适合于一般用途语料库的目标。其中,各类语料在时间段上的分配比例如表2-3所示。

表2-3 朗文语料库语料时间分段

时间	想象类	知识类
1900-1949	30%	10%
1950-1969	30%	20%
1970-目前	40%	70%

c 媒介：书面语语料库的来源包括：书本、报章和散页。散页主要指非出版物、广告、商业报告、政府告示或传单等。其中，想象类的来源主要是书本，知识类包括了书本、报章和散页。各种媒介所占的比例为：书本80%，报章13.3%，散页6.7%。

d 等级：这是保持分类一致性中最困难的特征。想象类文本的"高"级—文学小说，再次被确认为相对于知识类文本中相同等级的技术性文本具有更高的凸显性，原因是技术类非小说材料的读者范畴受限。

此外，朗文语料库还为文本设计了一些次要特征。

依据以上标准，朗文书面语语料库从2000多个文本资源中选取语料，其中600以上的文本来自于书籍，整个语料库共有2800万词次的语料可供研究者使用。

七、英国国家语料库BNC

在1991年至1995年间，英国国家语料库(The National British Corpus，缩写为BNC)无疑是全世界最具雄心的语料库编撰计划。该计划由政府出资一半，并由牛津大学出版社、朗文集团、钱伯斯(Chambers)出版社、英国国家图书馆、牛津大学和兰开斯特大学联合发起。这些单位分别贡献他们各自

在电子文本管理和出版、词典编纂和语料库分析方面的力量和经验,以便共同设计、开发和标注这个语料库。由于BNC的主题包含了口语和书面语文本,并且在规模上是有限的,使它有希望成为英国国家资源的主要参照点,犹如SEC、布朗和LOB语料库对第一代语料库研究作出的那种贡献,BNC的设计思想是良好的平衡,包含了来自书面语和口语的广泛体裁,并成为在教育、学术和商业用途上可以普遍使用的资源。

在BNC中共有4124个文本,90%来自书面语,10%来自口语。尽管在当时1000万词次的口语部分是英语口语语料库迄今最大的,里奇(1993)已经注意到BNC仍旧未能纠正绝大多数语料库中口语和书面语数据之间的严重不平衡现象。文献[38]是一个有关BNC的网站,下面关于BNC中语料分布的介绍。

1. BNC的书面语料库

BNC书面语语料库共包含3209个文本,每个文本都有各自的分类特征。在进行语料的选择时考查了文本的三个特征,它们分别是文本的发表时间、发表的媒体和主题领域。

(1)时间

表2-4 BNC文本的产生时间

时期	文本数	在书面语语料库中所占比例
1960-1974	53	1.65%
1975-1993	2596	80.89%
未标时间	560	17.45%

(2)媒体

在BNC语料库中,每个样本的大小都不超过4万词次。书面语语料库中样本的媒体来源及所占比例见表2-5所示。

表2-5　BNC书面语语料中媒体分布情况

媒体	样本数	在书面语语料库中所占比例
书本	1488	43.36%
报刊	1167	36.36%
散页(广告、传单等)	181	5.64%
非公开的散页(信函、随笔等)	245	7.63%
书面语化的口语	49	1.52%

(3)领域

目前,BNC书面语语料库中约有20%的文本是想象类文本,并且全部是1960年以后出版的;约有80%的文本是知识性的文本,全部都是1975年后出版的。想象类文本的样本数目可能高于采样周期内在全部出版物中小说所占的比例,这是考虑到文学与创作对社会所产生的持续不断的文化影响。表2-6给出了BNC语料库中文本的领域。

表2-6　BNC书面语语料中领域分布情况

领域	文本数	在书面语语料库中所占比例
想象类	625	19.47%
自然科学	144	4.48%
应用科学	364	11.34%
社会科学	510	15.89%
国际事务	453	14.11%
商业与金融	284	8.85%

（续表）

艺术	259	8.07%
信仰和思想	146	4.54%
休闲	374	11.65%
未分类	50	1.55%

2．BNC的口语语料库

BNC中约1000万词次口语语料，有两个主要来源：语境管辖材料(context-governed material)和统计采样文本。其中语境管辖类材料中共包含6154248词次，而统计采样的文本中共包含4122216词次。

(1)语境管辖的口语材料

为了达到文本类型的覆盖面，语境管辖材料包括了如下记录：讲课、辅导和教学等知识传授类事件；像演示、咨询和面试那样的事务性事件；像传道、政治演讲、公众会议和国会辩论那样的官方和公众的事件；像体育现场评论、俱乐部集会、广播电台的电话发言和闲聊那样的休闲事件。这些文本是在全英国12个地区系统收集的。表2－7给出其中各类事件的样本数和所占比例。

表2－7　BNC口语语料库的语境分布情况

语境管辖语料	文本数	在口语语料库中所占比例
教学类	144	18.89%
事物类	136	17.84%
机构类	241	31.62%
休闲类	187	24.54%
未分类	54	7.08%

(2)统计采样的口语材料

口语文本的第二种类型是由 124 个志愿者提供的、长达 2000 小时的录音转写材料，谈话包括生活的各个方面，系统地从英国的 38 个不同地区进行征集，他(她)们来自四个不同的社会经济人群，并且对年龄在 15 岁和 60 岁以上的男性和女性言谈者之间作出了恰当的平衡。每个志愿者都配备一台小型的携带式录音机，以尽可能不引人注意的方式记录他们两天内的全部对话。全体参与者都被告知他们的谈话已被录音，而且允许他们对录音带上的内容进行删除。对于对话情景和对话者的各种细节也都系统地予以记录。其中包括：背景、录音时的活动、处所、时间、日期、听众、自发程度、话题、参与者的性别、年龄、种族、职业、受教育程度、社会阶层、与对话者之间的关系、方言等。全部口语文本都经过正词法转写，带有停顿、犹豫、错误起头、重叠语音和重复的指示，以及像"大声喊叫或耳语"那样的超语言学特征。没有标注语音特征，而有少量的节律信息。因此，BNC 口语语料库还可以用来进行需要精细分析的语音研究。

3. BNC 语料库的服务

BNC 语料库中的每个文本都按照国际标准(SGML)进行了编码，语料库中的全部语料都使用了兰开斯特大学研制的词性标注系统 CLAW 进行了词性标注。BNC 还提供了强大的 BNC 语料库检索平台，能够完成对语料库的复杂检索。

八、国际英语语料库

在1988年,SEU的主席格林包姆(Greenbaum)提议研制大型的国际英语语料库(The International Corpus of English,简称ICE),目的是对讲英语的不同国家的英语进行对比研究,比较的范围包括书面语和口语。现在建立的国际英语语料库[39]包含20个平行的子语料库。每个子语料库的规模为100万词次,所有语料均从接受过正式教育、教育程度在中学以上、年龄超过18岁的成年人使用的材料中选取。这些子语料库既包括以英语作为第一语言或主要语言的国家,如英国、美国、加拿大、澳大利亚和新西兰等,也包括用英语作为官方语言之一或大多数人口讲英语的国家,如印度、尼日利亚、新加坡等。在这些子语料库中所有的语言材料都取自1990－1993年。虽然建立国际英语语料库的目的是进行英语的比较研究,但是其中的各个子语料库也可以独立用作对各国的英语进行语言描述的研究。通过使用国际英语语料库,研究者们可能发现哪些国家在英语的使用中用“different from(不同于)”而不用“different to”,还可能发现某些用法在不同地区的使用比例,如双重否定的使用等。

国际英语语料库建成的第一部分是英国的子语料库,该子语料库的组成见表2－8。之后,其他各国的子语料库几乎都采用了与之相同的结构。国际英语语料库的每个子语料库

都含有500个样本,每个文本的大小为2000词次。子语料库中只有40%的语料是书面语语料,300篇口语语料中大多数是公开的对话。

表2-8 国际英语语料库的结构

口语(300篇)	
对话(180)	独白(120篇)
私人(100)	不用稿子的(70)
直接谈话(90)	自发的言论(20)
远距离谈话(10)	不用稿子的演讲(30)
公共(80)	表演(10)
公开(80)	有讲稿的(50)
课堂授课(20)	广播新闻(20)
广播讨论(20)	广播谈论(20)
广播采访(10)	演讲(10)
国会辩论(10)	
法庭辩论(10)	
商业事务(10)	
书面语(200篇)	
非印刷的(50)	印刷的(150篇)
非专业的书面语(20)	信息类(学术的)(40)
学生的短文(20)	人文(10)
学生考试时的短文(10)	社会科学(10)
信函(30)	自然科学(10)
交际信件(15)	技术(10)
商务信件(15)	信息类(通俗的)(40)
	人文(10)
	社会科学(10)
	自然科学(10)

（续表）

	技术(10)
	信息类(报道)(20)
	新闻报道(20)
	教育(20)
	行政/正规教育(10)
	技能/爱好(10)
	劝导(10)
	新闻社论(10)
	想象类(20)
	小说/故事(20)

第五节　汉语语料库的建设

建立汉语语料库的要求最初是在汉语计量分析的需求下开始的。汉语计量分析包括汉语的字频、词频研究，以建立汉语的常用字表和常用词表。进行汉语的计量研究需要考察大量的语言事实，也就是要建立一定规模的语料库。最初这样的计量工作全部都由人工完成。在西方，第一部频率统计词典是德国语言学家凯定在 1898 年利用人工统计完成的。在我国，第一个进行现代意义的字频测定，是教育学家陈鹤琴在 1928 年完成的[40]。陈鹤琴和他的 9 名助理用手工操作，费

了两三年功夫，使用了六种文本类型共554498字次组成的语料库。统计结果表明，在这个语料库中共出现汉字字型4261个，在这些汉字中，出现300次以上的有569个字型，出现100次以上的有1193个字型。直到现在，陈鹤琴的统计数字还有很大的参考意义。直至70年代，我国仍采用人工方式完成过语料规模为两千一百多万字次的字频统计，这就是通常所称的“748工程”。

计算机汉字输入问题的解决，为汉语计算机语料库的研究奠定了物质基础。从20世纪70年代末开始，我国陆续建立了一批大规模的用于汉语计量分析研究的计算机语料库，《现代汉语字频统计》和《现代汉语频率词典》等工具书的出版就是这一阶段的典型成果。自80年代中期起，汉语信息处理界利用语料库主要进行了汉语自动分词系统的研究，以此作为机器翻译、自然语言人机接口和其他汉语信息处理的基础工程。到了90年代，随着计算机存储容量和计算速度的不断提高，汉字处理能力的不断增强，通过积极借鉴国外语料库建设的经验，我国大规模语料库的建设迅速进入一个新时期。最初建立的语料库都是为某一项专门用途而建立的，如字频和词频统计，而现在的语料库有着十分广泛的用途，不仅用于语言研究，还可以进行自然语言处理的研究等。本节介绍的是几个最典型的现代汉语语料库。

一、汉语词频统计语料库

1. 北京语言文化大学《汉语频率词典》项目专用语料库

北京语言文化大学语言研究所为编纂《现代汉语频率词典》[41],建立了一个规模为200万字次的现代汉语语料库。具体任务包括:采集各种题材和体裁的语言材料建立语料库,用人工方式对库存语料进行词切分,用计算机完成字频和词频统计。这项工程对现代汉语中汉字和词汇的实际使用情况进行了全面的调查研究,统计和分析了词语在各类语料中的分布情况、出现频率和使用度,对比它们的观察值和期望值;统计和分析汉字的出现频率和它们在词内不同位置上的构词能力。项目的目的是比较客观地从统计研究的角度展示出汉语词汇和汉字的使用概貌,区分出字、词的常用程度等级,对之进行随机抽样的覆盖率检验后,再对全部成果进行评审,最后按照预定要求编制完成了各种词表、字表和相应的统计材料。文献[42]详细介绍了他们的工作。

(1) 语料采样原则的确定

在建立语料库时,设计者根据语言在社会生活中的最广泛的应用情况,参照以前语料库的选材范围和内容,最后确定的语料选自报刊政论、科普作品、日常口语和文学作品等四个主题方面。同时又考虑到中小学语文课本大都选编了许多范文,课文一般使用合乎语言规范的字词,而且内容上能注意到

语言和文化知识逐步发展的特点,因此把我国1978－1980年全年通用的中小学语文课本也选作语料。这个语料库所选的语言材料内容及其分布情况如下:

Ⅰ类:政治、经济、哲学、法律、历史、地理、军事等反映现实社会生活的报刊文章及专著片断约44万余字次,占总语料的24.4%。

Ⅱ类:科普知识材料。选自具有中等文化水平的数理化、生物、医学、工程、技术、航海、宇航、科学史、科学家传记,以及与衣食住行有关的通俗科学文章约29万字次,占语料总量的15.8%。

Ⅲ类:日常生活口语材料。选自反映社会生活各个侧面的剧本名作(郭沫若、老舍、田汉、曹禺、吴祖光的作品)、相声、评书等。另外考虑到日常交往中口语的实际需要,还选用专题采录和随机采录的部分口语材料。两类合计约20万字次,约占语料总量的10.9%。

Ⅳ类:长、短篇小说、散文、童话等约89万字,占统计总量的48.7%。

在挑选文学作品时着眼于以下三个方面:1)以"五四"以来中国现代优秀文学作品为主,侧重在语言运用上较为成功的作品,兼顾各种不同的语言风格及流派。2)较多选用40至70年代作品。在题材上尽量照顾到各方面的内容和描写对象(战争、建设、工厂、农村、城镇、社会阶层、少数民族、历史人物、社会各个侧面),力求内容比例大体平衡。3)保持原著的

完整性。一万字次以内的短篇尽量选用全文,大部头著作则按原著章节截取相对完整并具有代表性的片段。

就全部语料来说,为使其中包括的词汇更全面和客观,对部分政论作品和科普书刊还采取了等距采样的方式选取供统计用的材料。剧本则抽选完整的幕次,只统计人物对话和独白,不统计场景及插叙的任务描述。中、小学语文课本,除成篇成段的文言文、诗歌、外语作品的译文外,全部都选作字、词频率统计用的材料。总之,设计者认为,语料的最佳采样原则必须兼顾到代表性、多样性和均匀性,总体数量的确定也必须科学地解决。量太少不能说明问题,量太多虽然能提高统计的精确性,但冗余度太大,不够经济。《现代汉语频率词典》的研制者在选定语料规模时充分注意到了这两个因素之间的权衡。

(2) 语料库的研究成果

汉语与使用拼音文字的外语不同,缺乏形态标记,词与词之间没有间隔标志,为汉语的词频统计造成了极大的困难。为此必须事先对语料进行词的切分。当时,汉语的自动分词研究大都只进行了方法不同、语料规模不大的实验,还不能说从理论到实践上已经解决"汉语自动切分"问题,有待改进和完善的地方还很多。这个语料库的建设项目就是在这种情况下于1979年开始的。当时是用人工完成了全部语料的词语切分,并且用人工标出每个样本的类型特征,最后再用计算机完成登记、统计、分析、综合等工作。

①利用这一语料库在现代汉语词汇方面完成的统计有：

a. 词的分类和综合，词次和词型数的总计和累计，计算相对词频和累计词频。

b. 计算每个词型的散布系数及使用度。

c. 根据词频高低编定词级，求出平均词长，计算各级词的词型数以及它们在各种词长中的分布。

d. 计算最高频词在分布中的观察值和期望值。

e. 求出单音节词、双音节词和三、四音节词及其他多音节词的百分比。

f. 按词的频次高低排列造表。

g. 按词的使用度大小依次造表。

h. 编出使用度在“5”以下，频次在“10”以下的低频词表。

②在汉字上完成的统计有：

a. 求出汉字字型总数，检索并生成汉字总表。

b. 统计每个字型的字次及累计总字次，计算每个字型的相对频率和累计频率。

c. 统计每个字型的构词条数，以及它在词内不同位置（词首、词间、词尾）上的构词条数。

d. 按字频高低排列编表，按构词条数多少排列编表。

2. 北京航空航天大学等单位的语料库

1981年11月19日，国家科委委托国家标准局下达了“现代汉语词频统计”任务，北京航空航天大学为主办单位，中国人民大学、北京大学、武汉大学等10个单位为参加单位。

该任务于1986年完成，1986年6月30日通过了“现代汉语词频统计”的国家级鉴定。下面是文献[42]中介绍的他们在这个课题下所做的工作。

（1）语料库的构成

现代汉语词频统计工程的选材范围为1919－1982年的正式出版物。从年代上分为四个时期：第一时期（1919－1949年），第二时期（1950－1965年），第三时期（1966－1976年），第四时期（1977－1982年）。每个时期都分为社会科学和自然科学两大主题类，每个大类又分为五个子类。具体学科的子类内容见表2－9。

表2－9 北京航空航天大学语料库中语料分布情况

类别		内容	字次数
社会科学	1	文体生活（包括服装、食谱、旅游、集邮等）	1147482
	2	历史、哲学（包括心理学、教育学、美学、社会学等）	2556804
	3	政治、经济（包括财贸、统计、管理等）	2747960
	4	新闻、报道（包括解放军报）	2880615
	5	文学、艺术（包括小说、散文、诗歌、戏剧、说唱文等）	4895724
自然科学	1	建筑、运输（包括邮政）	597046
	2	农、林、牧、副、渔	1211152
	3	轻工业（包括电子、日用化工、塑料、食品纺织等）	1448445
	4	重工业（包括矿山、冶金、机械、能源等）	1177586
	5	基础知识（包括数、理、化、生、天、地等）	2421802

选材的来源包括以下几个方面：

a. 报纸、期刊；

b 教材；

c. 专著;

d. 通俗读物(包括科普读物)。

以上的选材不包括翻译作品,着重在有代表性的作家及其他语言规范化的作品,但第一时期自然科学类的选材字数为零,这是因为在这一时期,国内几乎找不到有关自然科学的著作。

现代汉语词频统计工程的原始材料约三亿字次,通过随机和有规律(等距、分层)相结合的多层采样方法,采集样本的总规模约两千五百万字次。

(2) 语料库的使用

在建立语料库后,他们作了以下的工作:

a. 对1919年到1982年(分为四个时期)社会科学和自然科学(下分十个主题)的汉语材料,分时期、分主题进行了词频统计。同时,还对14种汉语标点符号进行了频度统计。

b. 首次实现了第一个实用的汉语自动分词系统CDWS。

c. 设计和实现了第一个完整的现代汉语词频统计软件系统。

d. 建立了一个有131161个词条的计算机词典。

e. 建立了一个有52个属性的汉字信息库。

f. 打印输出了1919-1949、1950-1965、1966-1976、1977-1982年等四个时期词语的综合频度,社会科学和自然科学的综合频度,以及总频度。7种打印结

果分别以汉语拼音和频度两种排序形式输出，共计打印纸1万余张。

这一工作在当时具有以下几个特点：

a. 是当时国内进行的一次规模最大、语料分布时间最长、分科最多的词频统计；

b. 语料采样分布合理，背景干扰小，统计精确度高；

c. 在国内首次实现了现代汉语计算机自动分词；

d. 采用“字词混合压缩码”，可以区分多音字，使统计结果更为精确。

以上两个现代汉语语料库都是专门为词频、字频统计而建立的，对词和字的使用进行了调查。很可惜，汉字编码标准的不统一，这些语料库在完成统计任务后没有能保留下来继续使用。以下各节介绍的语料库是几个比较著名的现代汉语通用语料库。

二、台湾中央研究院平衡语料库

台湾中央研究院平衡语料库(简称台湾研究院语料库，即Sinica Corpus)，是世界上第一个带有完整词类标记的汉语平衡语料库。该语料库的最终目标是建立五百万词次的汉语平衡语料库[43]。

1. 台湾研究院语料库的设计思想

台湾中央研究院词知识库小组，1990年前后开始致力于

中文语料库的收集[Huang & Chen 1992][44]，到目前已经收集近两千万字次的现代汉语语料及超过100万字次的古汉语语料[Huang & Chen 1994][45]。由于有了处理中文语料的经验，及处理大型电子词库的经验[陈克健等1996][46]，该小组已拥有足够的实力与人力来建设汉语的平衡语料库。他们的最初目标是要建立含两百万词次的语料库，几年后又将最终目标确定为五百万词次，接近计算语言学界通用语料库的规模。语料库的设计思想体现在以下三个方面。

(1) 遵循了台湾计算语言学会的分词标准

分词是中文自然语言处理的先决条件，但是由于中文在实际书写中词与词之间并没有空格断开，并且在确定词的概念上也存在争议。台湾计算语言学学会制定的分词标准在会员中通用，这样做不仅有助于资源共享，还可以不断取得语料库用户对分词结果的反馈，作为今后修改分词标准的依据。

(2) 采样时以自然段落为准，而不是以文章长度为准

布朗语料库的设计特色之一是为了求得语料分布上的平衡，每篇文章只随机抽取长度为2000词次的样本。词库小组认为，这样做将造成样本内容的不完整。此外，文章长短也往往是各种不同体裁的一个重要特征，如果人为地把文章截成长短一致的样本，则失去了这一特征。因此，台湾研究院语料库虽然仍然避免选取过长或过短的文章，但在选取文章后，随自然段截取样本。他们认为这样做可以得到较完整的语言信息的内容。

(3) 对语料采用多重分类原则进行分类

由于影响整个语言全貌的内在因素很多,单纯从某一特征如主题或体裁来界定平衡语料库是不够的。为了突破这种过于简单化的线性描述,词库小组对所有样本都给出了五个不同的特征值:体裁、题材、语式、主题和媒体。该小组目前虽然仍以主题为主轴进行语料的平衡,希望在有了更多的研究结果之后,可以同时利用多个特征来得到更加完善的平衡语料库。这样做的另一个好处是可以增加研究上的方便性。研究者可以任选某些特征的组合来建立适合自己的子语料库;也可以在子语料库上进行比较研究。

2. 语料的分类与选取

为了妥善管理以及选取平衡语料库的内容,研究院语料库中的每个样本前都有一定的标志来描述它们的体裁、题材、语式、主题、媒体、作者姓名、性别、国籍、出版单位等特征属性。

(1) 特征属性的指定

在参考了 LOB 语料库、布朗语料库和 COBUILD 语料库的管理经验之后,依据图书分类的一般原则,词库小组制定了一套分类中文语料的特征。这些特征用来说明每个样本的出处、写作方式以及谈论的内容等。主题反映了文本的内容,文类、文体和语式说明了文本的呈现形式。而出处则由媒体、作者、出版三个特征来标示,媒体说明了文本的出处来源。姓名、性别、国籍、母语表示了和作者有关的信息,出版单位、出

版地、出版日期和版次则记录了与出版有关的资料。

(2) 主题

主题依照文档内容和讨论重点来确定。研究院语料库大体上采用了图书馆的分类方法来制定主题的属性。

(3) 文类

文类说明文本的呈现方式,可以分为报道、评论、广告图文、信函、公告启事、小说故事、寓言、散文、传记日记、诗歌、语录、说明手册、剧本、会话、演讲、会议记录等。其中的语录都是来自报刊边缘的小语录,数量很少。信函约有三种来源,报刊杂志的读者来信、教科书内的书信范例和电子布告版里的书信往来。剧本都是来自小学课本,都是记叙文,主题为儿童文学,语式为书面化的口语。演讲包括三民主义的演讲稿,以及一些集成书或登于期刊中的演讲。

(4) 媒体

媒体根据资料来源分类,书面语和口语有不同的媒体来源。书面语的来源大体可以分为期刊、图书、书信、视听媒体、会议及其他;视听媒体包括了"女人,女人"电视节目的台词,还有一些电子布告版里的文章。电子布告版对大量语料库的建立极有帮助,可以不费时间而取得版权,也没有修改乱码和造字的麻烦,并且可以收集到多样化的文本。如果电子布告版内的文章表明了原来的出处,可以根据出处直接归类到其他媒体来源。其他表示的是不能归类于任何一种媒体的文档。期刊类分为报纸、学术期刊、一般杂志;图书分为教科书、

工具书、学术论著和一般图书。报纸包括中国时报、自由时报、儿童日报、中央研究院计算中心通讯。一般杂志包括天下杂志、光华杂志、海天游纵、翰林杂志、世界电影杂志；学术期刊包括医生简讯、民族所集刊。教科书有小学语文课本和师大国语中心提供的国语实用会话；工具书则收了词库小组的技术报告。学术论著是一些论文。一般图书包括了三民主义演讲稿、洪健全基金会的大众心理类书籍、时报出版的两本书等。口语语料来自大陆留美学生的日常对话等。

(5) 文体

文体是文本的写作方式，分为记叙、议论、说明、描写。记叙是将人、物的状态、性质、动作、变化等记录下来，一般记事叙述、信息报道的文章都属于记叙文。在所收集的文档中，记叙是最常用的写作方法。论说是提出自己的意见、主张以得到他人的认同、说服他人，一般评论的文章都是议论文。说明文的功用主要是分析事物的结构想象、道理、使人获得某些方面的知识和道理。所以仅以客观的文字说明事物的功能性质、形状等的文字属于说明文。描写是对人、物、事或景等做深刻的描绘，可能运用到比喻、修饰、排比、象征等多种描写技巧，来突出他的性质、特点，加深他人的印象。描写文包括抒发内心的抒情文章，如描述景物的游记或哈佛散文之类的也多半是散文。

(6) 语式

语式表示文本的呈现方式，仅以书面语或口语的方式表

达就大有不同。把语式分为 written、written-to-be-read、written-to-be-spoken、spoken、和 spoken-to-be-written。Written 即一般的书面语，也是语料库中收集最多的文本；written-to-be-spoken 是指剧本、台词等，为了让人在模拟显示会话情景下讲的，因为以事先演练过的方式表现出来，所以还是和实际中的口语不尽相同；spoken 即指一般的口语谈话，这类资料的整理较不容易，所以在语料库中所占比例很少。Spoken-to-be-written 是指会议记录之类的文本，由于还有修改整理的机会，可能去除了口语中的许多冗余部分，因此，值得另分一类，以区别于真正的口语或书面语。

3. 台湾研究院语料库的组成比例

台湾研究院语料库中语料分布主要以主题为主，目前定出的平衡语料库的内容比例为：

哲学：10%，	艺术：5%，
科学：10%，	生活：20%，
社会：35%，	文学：20%。

据此基准选取的语料来源及所占比例为(以字次为单位)：

(1) 报纸

中国时报：500566 字次，自由时报：1258334 字次，儿童时报：299260 字次，中央研究院计算中心通讯：957742 字次。

(2) 一般杂志

天下杂志：61994 字次，光华杂志：29840 字次，海天游

纵:138462字次,世界电影杂志:14869字次。

(3) 学术期刊

中央研究院民族所集刊:11225字次,中央研究院医生简讯:39507字次。

(4) 教科书

国民小学国语教科书12册:88744字次。

(5) 工具书

中央研究院资讯所词库小组的技术报告:6842字次。

(6) 学术论著、论文: 39076字次。

(7) 其他无法归入其他媒体的文本:8001字次。

(8) 图书

洪健全基金会的大众心理全书8本:216721字次,时报出版的巴西狂欢节:8011字次。

(9) 视听媒体

台湾学术线路里刊登的文章103955字次。

(10) 会话访谈:28831次。

4. 语料库的分词和词类标注

台湾研究院语料库进行了自动分词处理。切分的过程是:以台湾中研院词典的九万词条为基础,将语料切分成一个个独立的词。没有列在词典中的成分,则以字为单位,被一一切开。然后,以构词规律对衍生性强的词缀及数字组合成分进行结合成词的工作。在切分时,分词的大原则为:(1)有独立意义的语法类可依类分为一分词单位。(2)惯用的语言成

分依人的使用习惯切分。(3)语义失去组合性,或语法起变化,合为一个分词单位。(4)有明显的分割标记时需切分。(5)同形异构的成分依实际语境切分。(6)原则有冲突时由计算语言学会统筹协议解决。据此,台湾研究院语料库中的全部语料都进行了分词处理。

采用先机器后人工校对的方法,对全部语料库进行了词类标注,在标注时采用的词类标记集共含 46 个标记。除了标记词类外,还对语料库中的某些特殊句法表现做了标记,使用的特殊标记共 8 个,是针对动补式动词和动宾式动词的可拆现象、合并词中插现象、名物化结构和外来语设计的。

此外,台湾研究院语料库还提供了对语料库进行关键词检索的功能,可以从语料库中找出含有某个特定词条的上下文,并给出他们的分词和词类信息,供语料库的使用者查阅。表 2－10 和表 2－11 分别给出没有词类信息和含有词类信息的两种查询结果。

[显示词类及特征]

[至主画面][再处理]

表 2－10　　**共 19 笔资料**

:"既然你非和我比剑不可,我也乐于试试你的本事,不过,我随身带着主君的

且又影响家庭生活其巨,所以没有出去试试闯天下,老有遗憾。此时,不妨以

用过一次没什么感觉了下次再去买一副试试只是不便宜耶明天还

要考动力学我要去。

对非物理学者而言你可试试 Close Sutton 和 Marten 的此书包含了许多

的绝世刀法。学了两年,懒残大师有意试试他的功力,便把他叫来禅室,其时外面

还在纽约的话,我们就去买双冰刀鞋来试试,你就说在屏东溜冰的故事,穿好那种

带动人际之间的热络气氛,你是否也想试试,以镖会友一番? 快加入飞镖行列吧!

自认年轻貌美,身材还过得去,不妨去试试运气。假设你很不幸地没有上述这些

相信的迷信疗法,她也抱着一线希望去试试,两年来她为了能传宗接代做个真正的

及钢琴等自动演奏乐器,游客也可试试身手,弹奏乐曲,体验开演奏会的

了,你走吧! 狐狸说:象啊! 老虎只是试试你的胆子大不大而已,没想到你的胆子,

真是一举两得呀! 各位有空时不防试试我的消暑妙方,不然,你总不能一箭

为什么它的名字中有个虿字,大家不妨试试身手,看能不能徒手抓到它们。图说:1

是不喝酒的,看看杯子这么可爱,也想试试。"格林哥说:"傻瓜,这不是杯子,

中国人上了床,功夫特别好,不信可以试试。"她们信以为真,极感兴趣地打量着

一句:"不论你同不同意,今夜我要试试。""喔! 中国人! 难道你们连做爱都

那人关了门,开始剥身上衣裳:你看我试试夹袄合适不。直到那色鬼剥光衣裳露出

一些创意，不是一举数得吗？大家不妨试试。大台北经济证券，新台币汇率昨日再，

他刚学会开车，回国后，喜欢在台湾试试他的开车技术。没有开两条街，就大叫

表2-11　　共19笔资料

不可(D)，我(Nh)也(D)乐于(VL)试试(VF)你(Nh)的(De)本事(Na)，不过(Cbb)

巨(VH)，所以(Cbb)没有(D)出去(VA)试试(VF)闯(VC)天下(Nc)，老(D)有(V-2)

再(D)去(D)买(VC)一(Neu)副(Nf)试试(VF)只是(D)不(D)便宜(VH)耶(T)

学者(Na)而(Cbb)言(VE)你(Nh)可(D)试试(VF)Close(FW)Sutton(FW)和(Caa)

年(Nf)，懒残(Nb)大师(Na)有意(VL)试试(VF)他(Nh)的(De)功力(Na)，便(D)把(P)

买(VC)双(Nf)冰刀(Na)鞋(Na)来(D)试试(VF)，你(Nh)就(D)说(VE)在(P)屏东(Nc)

你(Nh)是否(D)也(D)想(VE)试试(VF)，以(P)镖(Na)会(VC)友(Na)一(Neu)

还(Dfa)过得去(VH)，不妨(D)去(D)试试(VF)运气(Na)。假设(VE)你(Nh)很(Dfa)

着(Di)一(Neu)线(Na)希望(Na)去(D)试试(VF)，两(Neu)年(Nf)来(Ng)她(Nh)为(P)

乐器(Na)，游客(Na)也(D)可(D)试试(VF)身手(Na)，弹奏(VC)乐曲(Na)

象(Na)啊(T)！老虎(Na)只是(D)试试(VF)你(Nh)的(De)胆子(Na)大(VH)不(D)

位(Nf)有空(VH)时(Ng)不妨(D)试试(VF)我(Nh)的(De)消暑

(VA)妙方(Na),

蚤(Na)字(Na),大家(Nh)不妨(D)试试(VF)身手(Na),看(VE)能不能(D)徒手(D)

这么(D)可爱(VH),也(D)想(VE)试试(VF)。"格林哥(Nb)说(VE):"傻瓜(Na)

好(VH),不(D)信(VK)可以(D)试试(VF)。"她们(Nh)信以为真(VH),

今夜(Nd)我(Nh)要(D)试试(VF)。""喂(I)! 中国人(Na)! 难道(D)

衣裳(Na):你(Nh)看(VE)我(Nh)试试(VF)夹袄(Na)合适(VH)不(T)。直到(P)

吗(T)? 大家(Nh)不妨(D)试试(VF)。大台北(Nc)经济(Na)证券(Na)

后(Ng),喜欢(VK)在(P)台湾(Nc)试试(VF)他(Nh)的(De)开(VC)车(Na)技术(Na)

三、中文五地区共时语料库

中文五地区共时语料库(Linguistic Variety in Chinese Communities(简称 LIVAC 语料库))是由香港城市大学开发的。它采用共时同步手法收集语料,目的是适量地选取具有代表性的各地语料,对汉语在各华语地区的实际用语和用法,以语料为基础作出全面的描写和分析,并据此提出和加强一些具有解释性的理论。这样做还有利于推进中文词汇语法的研究。文献[47]详细介绍了香港城市大学语言资讯研究开发研制的这个语料库。

1. LIVAC语料库的特点

该语料库的特点主要体现在以下两个方面：

(1) 像LIVAC语料库这样同时在五大华语地区进行大规模系统的语料采集，是史无前例的。这五个华语地区除了中国大陆外，还包括了香港、台湾、新加坡和澳门。

(2) 设计并坚持在相当长时期中采集"共时性"的语料。"共时"语言学的本质是研究时间历程中某个假设点上的语言，描写当时的语言的"状态"，而大多不顾各状态的前例和以后的可能变化。语言犹如有机的生命体，千变万化，这种情况以新词的产生和扩散，已有词的变化或消亡最为显著。因此，在某个假设点上收集语料，也最好能容许某种程度的历时性，以便观察某些词语的扩散和变化情况，包括其可能在不同地区产生的不同影响。因此，LIVAC语料库把语料采集的第一阶段定为三年。为了能取得合适的语料，LIVAC所采用的报纸语料，都分别是各地区在同一天出版，而主题尽可能相互对应的部分。LIVAC所谓的"共时"，一方面比语言学上的共时更为严格，而另一方面则是"共时"之中兼顾若干历时性，目的是希望有机会能了解某些类别词语的整个成长或消亡过程，及其由来与意义。

2. LIVAC语料的范围和采样

LIVAC语料库从1991年开始孕育，1993年获得初次资助后开始策划，具体步骤是，在香港、澳门、上海、新加坡和台湾五地，每四日选定一天的报纸。资料的内容包括社论、第一

版的全部新闻和文章、国际和地方版以及特写和评论。每天每地收集的数量为两万字次。从 1995 年 7 月至 1997 年 6 月的两年期间，LIVAC 语料库每年所得的语料总数分别为 67801 万字次和 84543 万字次。

3. 语料的自动切分和语料库的建立

建立 LIVAC 语料库的主要目标是对词语进行分析，所以首要任务便是对文本进行词切分，LIVAC 语料库采用最大匹配的分词方法，首先由计算机自动分词，并且在切分过程中对一些特别类型的词语(如数词、人名、地名等)自动加上一些标志，以方便以后分析和使用。自动词切分的正确率在 95% 以上。然后，由人工对切分结果进行校对，主要有两个步骤：首先对整篇已经经过分词的文章进行检查和修正，然后自动抽出所有的词语，第二次人工检查是否有不合理的词语存在，最后把这些词语装入 LIVAC 词典中，以进一步提高分词程序的准确度。

人工校对后的文章，通过计算机自动阅读，取出其中的信息建立 LIVAC 语料库。除了将所有的词语登记，还要记录来源地区及日期资料，并且加上其他有用项目，如汉语拼音和粤语拼音。为方便对所收录语料进行检索，所有新闻的标题及内容，都会被存放在文本语料中。此外，为提高对大量文章进行检索的速度，同时建立文本的索引文件，记录每一个词语在不同地区、日期、段落、句子，以及在语句中出现位置等。从而建立起一个快速而全面的检索系统。

LIVAC检索系统提供了如下检索功能：

(1) 词语检索

显示词的属性，包括该词的汉语拼音、粤语拼音及其所有的英文解释；显示统计数据，即该词在各地区出现的次数。列举对比词，即显示该词在各地区的对比词。

(2) 文本检索

允许用户以一个词语或词语式样、词类、汉语拼音或粤语拼音，在配合其他相关条件，从语料中检索出适当的语句。可以实现四项检索目标：列举目标词及其前后词语，列举包含目标词的短语；列举包含目标词的句子；列举包含目标词的文本标题。

四、现代汉语研究语料库

现代汉语语料库被确定为国家教委"八五"人文、社会科学规划项目，同时被确定为北京语言文化大学"八五"科研规划重点项目。现代汉语语料库的研制目的是，为从事现代汉语研究、汉语教学以及汉语信息处理研究工作者提供一个以大规模真实文本为基础的研究平台，促进汉语研究的深入开展[48]。

1. 语料的选择与采样

现代汉语研究语料库有两个层次：第一级是规模达2000万字次的粗语料库，第二级是一个规模为200万字次的精语

料库,即经过分词和词性标注的语料库。相应的有两级采样。

第一级采样是从约6000万字次的素材中抽取出2000万字次的粗语料。采样的原则主要从文本的完整性,文本长度等方面考虑的,比如,1000字次以下的文本一般不抽,文本不完整的不抽。6000万字次的素材包括中新社《中国新闻》1992－1993年共2000万字次,新华社1993年的电讯稿1500万字次,《人民日报》1994年全年近2000万字次。另外从馆藏图书中抽选文学类、口语类共250万字次语料人工录入。从这些素材中建立的2000万字次的粗语料库构成情况如下:

《人民日报》(1994年全年)　　1000万字次

《中国新闻》(1992－1993年)　　500万字次

科普等类书籍　　250万字次

文学作品(录入样本)　　150万字次

(其中小说100万字次,散文30万字次,报告文学20万字次)

准口语材料(录入样本)　　100万字次

(其中对话〈话剧剧本〉60万字次,独白〈单口相声、演讲词、对话、故事等〉40万字次)

从以上数据看出,在一级语料库中新闻语料所占的比例为75%,科普书籍为12.5%,文学作品为7.5%,准口语材料为5%。

第二级抽样是从不含《中国新闻》和科普等类书籍的1250万字次语料库中按照设定的比例随机抽取的200万字

次样本。考虑到《人民日报》是一份综合性的报纸，题材丰富，而另外750万字次语料在题材、体裁上比较单一，而且这些题材在《人民日报》的语料中也占相当的比例。在设定比例时，同时考虑主题和体裁两个方面，分别从这两个方面给出文本的分类体系。在具体考虑各类的比例时，基本上坚持了既要全面又要有重点的原则。例如，从主题方面来说，政治、经济、文学三类的比例很高，而历史、地理、军事等的比例就比较低。从体裁方面来说，记叙、议论占了绝大多数，说明和应用的比例就很低。具体的采样步骤是：(1)建立文本属性库。文本属性包括文件名、字数、出处(书刊名、出版单位)、出版时间、主题类别、体裁类别等。(2)设定语料分布。(3)由程序随机抽样。二级抽样的结果见下面的表2－12、表2－13和表2－14。

表2－12　书面语语料的题材分布

主题类	比例	字数(万字次)	篇数
政治和法律	15%	30	140
经济	15%	30	129
文学	18.75%	37.5	113
文化教育	7.5%	15	83
社会生活	7.5%	15	80
科技科普	6%	12	69
体育	4%	8	19
地理旅游	2.5%	5	34
历史考古	1.25%	2.5	13
军事	2.5%	5	32

表2-13 书面语语料的体裁分布

体裁类	比例	字数（万字次）
小说	10%	20
散文	10%	20
报道	25%	50
报告文学	5%	10
回忆录	2%	4
论文	9%	18
评论	14%	28
知识小品	0.5%	1
说明文	1%	2
提要	1%	2
公文	1%	2
事务问题	1%	2
书信	0.5%	1

表2-14 口语语料的体裁分布表

体裁分类	字数(万字次)	篇数
话剧	25.5	65
单口相声	9	28
评书	1.2	2
演讲 讲话	3.8	31
故事	3.5	26

对二级语料在加工前都进行了预处理。首先过滤文本，比如过滤掉话剧剧本中的非对话部分，还有其他样本中的一些版面信息，然后在每个样本上加段标记，按照主题属性分类重新命名。

2．语料库的加工

汉语语料库的标注，首先要对语料进行分词。为了使分词工作有一个比较可靠的依据，保证分词的结果有较高的分词一致性，制定了可操作性强、高度具体化的分词规范。选取有代表性的20万字次的现代汉语语料，找出其中所有的双音节和三音节组合，然后对这些组合进行多角度的分析，其中包括内部结构（组成成分是否单用、内部结构关系、组成成分的功能类、内部能否扩展等）、整体功能类、语义组合情况、音节构造、语体因素等各个方面。此基础上，来检验现有分词理论和方法的有效性，最后制定出分词细则。分词原则主要体现在三方面：(1)词是一个句法语义范畴。(2)词的划分不是绝对的。(3)应该区分语料中的不同层次。之后，确定了判断单语素是否成词的标准和判断多语素组合体成词的标准。

对现代汉语研究语料库中的二级语料库都进行了词性标注。在参考了汉语语法学界的研究成果，以及信息处理界有关语料库研究中词性标注的成果，指定了在确定词类标记的两个主要原则，一是完全根据词在句子中的句法功能确定词性，二是词的分类粗细适度。为此，采用了多级的词类的分类体系。如名词第一级的标记为“n”，第二级又分为专有名词、普通名词、时间词、处所词、方位词等五类。共设计了含有85个词类标记的标记集。语料的标注采用了CCID提供的分词和词性标注一体化的标注工具。机器标注完后，人工对200万字次的语料进行了逐词校对。

为了使现代汉语研究语料库为语言研究者服务，开发了

一个具有建库、检索、统计等功能的应用平台。利用一级语料库可以进行基于字串的检索；利用二级语料库可以进行基于词和词性的多种检索和统计，比如可以检索包含具有某个指定词性的词的句子，也可以检索某个短语格式，并且还提供了多种形式的输出形式。

五、汉语的精加工语料库

汉语精加工语料库是国家自然科学基金重点项目——"语料库语言学研究的理论、方法和工具"的一个子课题。该子课题的目的是建立一个汉语语料库的精加工体系，该体系主要包括：《现代汉语语料库文本分词规范》、《现代汉语语料库词性标注规范》；语料库选样的基本原则及语料分布；最后得到一个200万字次(不包括标点符号)的经过分词、带有词性标注和部分语法信息的、分布合理的均衡语料库，以及1.2亿词次的质量可靠的生语料库。

该语料库取得的成果可以形成现代汉语词汇、句法研究的支撑平台，对中文信息处理、汉语语言学研究、汉语教学研究等具有重要意义。

1. 语料库选材

(1)选择语料的原则

语料的选择主要遵循以下几条原则

a. 主要采用90年代的现代汉语语料(小部分为80年代

的语料),能够反映当代汉语的最新面貌。

b. 语料的采集以篇为单位,一般选取全文以保留篇章信息。

c. 文体分布为主线,领域分布为辅线,文体分布优先于领域分布。

d. 语料仅限于书面语以及表义连贯明确并能用书面语转述的准口语材料(主要来源于剧本、谈话录、演讲录等)。文学语料占较大比例,适当增加准口语和应用文体语料,以避免过于书面化。

e. 不包含方言及港、澳、台地区的出版物。

(2)语料的具体分布

语料按文体分为文学、新闻、学术、应用四类,总规模是200万汉字。各类语料分类说明如下:

a. 文学类

1) 小说(包括普通小说、爱情小说、科幻小说、侦探小说等门类)

2) 散文(包括杂文、小品文等)

3) 回忆录(包括传记)

4) 报告文学

5) 剧本(包括谈话、演讲)

b. 新闻类

1) 新闻报导(包括政治、经济、军事、工业、农业、商业、科技、教育、体育等各领域)

2）社论及述评

3）社会生活、休闲体（包括旅游、烹饪、服装、园艺、嗜好等）

c. 学术专著（包括人文与社会科学、自然科学的各子类）

d. 应用文体（包括告示、通知、信函、报告、合同、备忘录、产品说明书、广告等非正式出版物）。

语料的具体分布情况参见表2－15和表2－16。

表2－15 语料分布统计

分类	篇数	字次数	比例	标点符号数	词次数	比例
文学	295	880057	44%	148453	760337	48%
新闻	376	600490	30%	86163	438095	28%
学术	29	402623	20%	52823	278728	18%
应用文	258	119488	6%	28727	91929	6%
合计	958	2002658	100%	316116	1569089	100%

表2－16 文学语料分布统计

分类	篇数	字次数	百分比	标点符号数	词次数
小说	199	648796	32.5%	112749	566730
散文	37	80067	4%	10347	65453
回忆录	29	50401	2.5%	6908	38338
报告文学	13	50019	2.5%	8225	40386
剧本	17	50774	2.5%	10224	49430
合计	295	880057	44%	148453	760337

2. 分词规范

在制订分词规范时，特别注意了以下两点：

(1)规范是在对分词进行大量研究的基础上制定的。

a.吸收以前的成果。充分借鉴计算语言学和汉语研究的

成果,尽量和已经颁布的国家标准取得一致。但是又体现出自己的特色。

b.对切分有较大分歧的部分进行专门研究。曾从经过7位在读硕士研究生人工校对的150万字的语料中抽取了71万字,统计出分合意见不一致的切分单位。并对这些切分不一致的单位进行分析。从研究结果来看,词和短语界限不清的情况主要集中在双音节和三音节结构中;从语法功能上来看,主要集中在名词性和动词性的结构中。

c.一边使用一边修改,使规范的准确性更高,涵盖面更广。规范的制定不是盲目的硬性规定,一切以实际语料的分布为基础。

(2)采取必要措施,保证分词的一致性。

a.对一些长期有意见分歧的语言单位进行了硬性规定。例如,对2至4音节"名+名"和"动+名"的偏正结构,若其中一部分音节长度为1,则整体不切分。

b.规范中很多细目是专门为一个词条或几个词条而列的。

例如:"半"单独作数词用时切出:

半/斤、一/斤/半

但是以下包含"半"的组合不切分:

一半儿、多半儿、两半儿、大半儿、一多半儿、一大半儿。

3. 词性标注规范

制订的词性标注规范中共含标记119个，其中词性标记95个，标点符号标记24个。这95个词性标记采用层级体系，最多有三个层级，如“npf”中“n”是最高层，表示名词，“p”是第二层，表示名词中的专有名词，“f”是第三层，表示专名中除日本人名及朝鲜、越南的汉式姓名之外的外国人名。最高层有22个大类：名词(n)、动词(v)、形容词(a)、状态词(z)、区别词(b)、时间词(t)、处所词(s)、方位词(f)、数词(m)、量词(q)、副词(d)、代词(r)、象声词(o)、叹词(e)、连词(c)、介词(p)、助词(u)、语气词(y)、插入语(l)、成语(i)、词缀(k)、阿拉伯数字、英文等(x)。

词性标注规范具有以下特点：

(1)充分注意词性标注难点。当给语料库中的词赋予词性标记时，大致会遇到以下几种情况：

a. 意义固定，语法功能单一的词，这类词不会有兼类的问题，比较容易归类；

b. 具有明显不同的语法功能，意义上有很大差别的兼类词(包括同形词)，这类词应该归入两个或多个类中；

c. 有一部分词语，虽然句法功能有所不同，但是要不要看成兼类意见还不一致；

d. 语法功能和意义方面可以归入两个或多个类中，但是实际又不能归入两个类中，比如形容词和不及物动词；

e. 语法功能不典型，找不到合适的类；

f. 其他类的情况，包括分词带来的问题、临时用法等。

词性标注的难点主要集中在后四类情况中，如形容词和不及物动词的界定、区别词和副词、名词、动词、形容词等的界定问题。在制定标注规范时对这些类型给予了充分注意。

(2)部分词类采用层级标记，为某些专门研究提取词类标记信息提供了方便，可以不仅研究一个大类，也可以专门研究某一小类。

(3)专有名词的细划，尤其是对指人的专名划分细致，为专名的识别和测试提供丰富的信息。

(4)标注了部分句法信息。主要集中在动词部分(形容词带宾语时也标作动词)。这样可以对和动词相关的一些短语进行专门研究，从动词带宾语角度，可以对一些特殊句式进行研究。

(5)对一部分频度比较高的词语给以专门的标记，以便于进行专门的研究。

4. 汉语精加工语料库的质量保证

语料库的分词和词性标注加工过程是在机器自动切分和标注的基础上进行人工校对，人工校对是语料库加工过程中最关键的一环。语料库的人工校对过程中包括分词校对和词性标记校对两个方面。分词中最容易出现的问题是同一单位分合不一致，词性标注中最容易出现的问题是词形、语法功能和意义完全相同的单位的标记不一致。根据它们各自的特点，我们采用了分阶段进行校对，而且，不同的阶段采用不同的方法进行一致性检查。

(1)分词校对

分词校对分两个步骤:第一步采用逐篇逐词校对;第二步则编写分词一致性检查程序,将可能出现的分歧全部找出,形成切分不一致词表,进行词表驱动的校对。

(2)词性标记校对

词性标记校对也同样分成两个步骤,第一步采用逐篇逐词校对。第二步因为分词不一致问题已经在分词校对过程中解决了,所以我们只需要生成一个包含全部词语和标记的词表,进行词性标记的词表驱动的校对。这样就避免了同一单位不同标记这样的不一致。

该语料库已准备放入网站,提供给更多的研究者使用。

第3章 语料库的加工和管理技术

第一节 语料的索引及其应用

语料库建立之后，将提供给各种不同的研究者使用，使他们能够访问语料库的内容，对现实的语言现象进行分析。因此，语料库应该向用户提供的最基本的工具就是语料的检索工具。

一、逐词索引

计算机语料库一般以逐词索引(concordance)方式向用户提供指定词形在语料库中每次出现的相关信息。逐词索引程序记录了每个词形在语料库中每次出现时的位置，据此就可

以提供指定词形每次出现的上下文信息。这样的信息既可以直接显示在计算机屏幕上，也可以保存在某一个文件中。保存起来的这个文件就称为逐词索引文件(concordance file)[49]。

在进行逐词索引之前，需要为语料库中的每个词建立一个索引，记录这个词在文本中每次出现时的位置，然后将建立的索引数据文件按一定顺序重新分类排列，以便查找。如可以词的拼音字母顺序建立语料库的索引表。这样在进行逐词索引时，便能够快速地找到指定词形的每次出现及其上下文信息。

逐词索引的最简形式就是位置索引。它指明任意一个词形在文本中每次出现的位置，同时也可以提供该词形在语料库中的出现频率。另一种是以(文本)行的形式提供索引，直接提供一个词形每次出现时的上下文信息。

1. 关键词居中索引

最有用也是最常用的索引形式就是关键词居中索引(Key Word in Context，简称 KWIC)。被检索的关键词在每个索引行的中央出现，关键词两边至少各有一个空格，空格两边各有一段可以指定长度的上下文。在这种显示方式中，关键词自上而下形成整齐的竖列，安排十分醒目。图 3－1 给出了一个关键词"is"的前后各四个词的关键词居中索引。

of activity and communication is only one of them
communication where the activity is halted in time if

whole process the activity	is	obvious enough the nervous
reader in his armchair	is	making continuous fast and
radio listener his brain	is	highly active if he
human communication through language	is	only a small sub-section

图3-1:对"is"进行KWIC的部分检索结果

关键词左右的上下文长度可以根据需要任意设定。如±6表示该索引行的上下文为关键词向左或向右各6个词。有些索引工具还可以将上下文扩展到关键词所在的一个完整的句子或段落。

2. 索引行的排序

索引行在逐词索引文件中的排列可以有多种形式。经常使用的顺序是:按索引行在语料库中出现的顺序存放,或按索引行的字母顺序排列。例如,按关键词右边第一个词的字母顺序排列该关键词的同现行。这种排法突出了中心词引起的词组。另一种排序方法是按关键词左边最后一个词的字母顺序排列。当关键词是动词时,这种方法有时能够帮助用户迅速找出动词的主语,从而为了解主谓搭配和篇章的主题提供有用线索。我们还可以按照左、右词出现的频率从高到低排列,把与关键词最常同现的词形首先集中存放在一起。这种表现形式对词语搭配研究极其有用。

对语料库中的高频词,还可以用采样索引的方法来减小索引样本的规模。例如,可以在检索某词时将参数定为每10行索引一处,从而将该词的索引样本缩小到原来的十分之一。

在英语中,利用统配符(*)还能查找与某一词的原形有

关的其他词形。如查找“photo＊”，就能检索出 photo，photograph，photographer，photography 等有关的词形及上下文同现行。也可以用统配符来辨识并索引某些特殊词形和语法特征。例如，用“＊ing”能列出所有含“ing”的词形，用“＊?”能找到所有的直接问句等。

利用组合逻辑和“with/n”操作符，还能检索出与词组、短语以及与关键词有关但被其他词隔开的上下文。

二、逐词索引在语料库中的应用

借助于计算机强大的计算与信息处理能力，用户通过逐词索引软件可以快速地从语料库中检索到所需要的内容，并可以将它们与所在的语言环境一起观察，或与相关的语言现象对比研究。语料库的这种应用具有巨大的价值，在文学、语言学、语言教学和自然语言处理系统等研究中得到了越来越广泛的应用，并产生了丰硕的成果。以下列举几个方面的应用[23]。

1. 在文学研究方面的应用

建立文学名著语料库，并且建立起它们的索引文件。这些索引文件被广泛地用于文学领域的研究，是考察某一位作家或诗人的作品特色、词汇使用特点和语言风格的理想工具。

建立文学作品的索引文件，也可以被语言学家用来研究特定语体的特点。例如，通过建立莎士比亚戏剧的索引，可以

研究英语中第二人称代词的多种形式(ye, you, your, yours, thou, thee, thy, thine)的使用场合,从而提出并验证社会语言学方面的某些假设:谁是讲话者,对话者,对话的场合是什么等。又如,从歌德作品的逐词索引文件中,可以通过研究一些词语的使用,管窥歌德作品语言在过去200年中的发展变化趋势。像knabe一词,在歌德的作品中比现代作品中用得普遍得多。所有格的使用也是如此。可以通过与现代作品的对比,对这些变化进行量化的分析。用逐词索引还可以对文学作品中的许多侧面的词语、类型与结构进行研究,认识同一作家早期与晚期作品的异同,以及不同作家文风的异同。甚至可以为判断某部作品的著作权的归属与文学作品的分期断代提供有力的依据。

2. 在语言学研究方面

里奇(Leech 1992)曾指出,以计算机为基础进行的最简单、最有用和最普遍的工具就是逐词索引。典型的范例是:

(1)词汇学研究:辨明某词项在指定上下文中的含义,以及与该词用法有关的句法、文体和语用特征等。

(2)归纳性语法研究:辨识并归纳出某种语法项或结构体在句法、语义、语用、语体方面的典型例子。带有词类和句法标注的语料库更适合这一目的。

(3)在语言教学方面:在英语教材编写上,比伯等(1994)曾对"名词+后置修饰语"的结构进行了研究[50]。他们查阅了80年代多本英语语法书中对这一结构的解释,发现各书均

把这种关系作为重点,讨论的篇幅合计达 60 页。而对“名词+介词短语做修饰语”的结构讨论篇幅却较少,篇幅合计不足 5 页。

用逐词索引软件对 LOB 语料库和一个含有 11.5 万词次的私人信函库的索引却表明,“名词+介词短语做修饰语”的结构,比“名词+从句做修饰语”的结构的使用频率高得多:每千词中二者的使用频率是 23.3:5.5。而且已有研究表明,“名词+介词短语”的结构对外国学生在学习英语时是一种困难结构。

这说明教学语法对语言现象处理的轻重缓急,常常不符合语言的实际使用情况。在编写这类教材时,不但应该考虑语言现象的难易程度和教学方法,而且应当把它在实际应用中的情况考虑进去。在这一方面,Collins COBUILD English Grammar 就是在参考了 COBUILD 语料库中大量实际语料的基础上编写的。该书利用索引技术提供了大量语法结构及其实用例句。

第二节　语料库语言学中的统计

语料库是计量语言分析的重要资源。但是在语料库语言学中,统计量的使用并不只是对语料进行简单的计数。这里使用的一些统计技术,不但可以对复杂数据进行数学分析,从

无序的数据中找出它们的规律，还可以用来说明文体、写作风格和语言之间的差异。

这一节简单介绍文献[2]中提出的在语料库语言学中最有价值的统计方法。这部分只是集中介绍在语料库语言学中一些最重要和最常用的方法，不可能罗列所有的方法；而且还试图简单地说明这些统计技术在语料库语言学中的应用，比如如何使用，有何特殊意义，而不提供它们的实现细节。

一、频率计数

计数是对语料库进行的一项最简单的统计工作，也就是根据某种特定的分类对文本中指定条目的每次出现进行计数。计数得到的各个条目在语料中的出现次数，就是这个条目在语料整体中的出现频次。在英语中，这个条目可以是词形，去掉词的屈折变化后的词型，或是词类等，而在汉语中这个条目可以是字或词。如果没有特别说明，本节采用的分类都是指词形。计数过程如下：从前到后依次审视语料库中的每个词，如果该词形以前已经出现过，则对该词形的出现次数加1；否则将该词形添加到频次表中，并且置该词形的出现次数为1。

用词形的计数来说明该词形在整个语料库中出现的频次，在语料库语言学中有重要用途。如果按词形在语料库中的出现顺序排列得到的频次表，可以研究词语在文本中的分

布情况。比如,一篇技术报告,如果在一段时间里很少出现技术词汇,而后又突然出现很多技术词汇,这个特性可以作为切分文本段落边界的一个线索。即可能是一段概要的结束,也可能是非专业人员对一个技术主题的引言;把频次表中的词形按字母顺序排列,主要用途是建立语料库索引,这样可以加快语料库的查询速度;把频次表中的词形按它们在语料库中出现次数的高低降序排列,有助于对文本与词汇的研究;把一个特定文本的频次表与另一个在较大语料范围内统计得到的频次表加以比较,位于两个频次表最前面的那些词形是经常在语料库中出现的词形,一般说来它们是稳定的。去掉这些高频词后,可以帮助推测这个特定文本的关键词。

二、比例

一个条目在语料库中所占的比例,是指这个条目在语料中出现的频率除以语料中所有条目出现的总次数。虽然频率计数是对语料库中数据进行量化处理的有效方法,也是基于语料库研究中经常使用的方法,但是这种方法也存在着某些不足。比如,在比较两个不同的数据集时,就暴露出这种方法的缺点了。现在假设要比较英语的口语语料库和英语的书面语语料库,这两个语料库的频次表中分别记录着各个词形在各自语料库中的出现次数。当两个语料库的规模不同时,用这个频次表很难对这两种语言进行对比。虽然某个词形在一

个语料库中的出现次数大于它在另一个语料库中的出现次数，但是有可能它在前一个语料库中的所占的比例却小于在后一个语料库中的比例。假定所比较的英语口语语料库和书面语语料库的规模分别为5万词次和50万词次，词“boot”在这两个语料库中的出现次数分别是50次和500次。从次数上看似乎“boot”在书面语语料库中比在口语语料库中更常出现，但情况并非如此。现在让我们分别计算“boot”的词次在这两个语料库中所占的比例：

口语：50/50000＝0.1％，

书面语：500/500000＝0.1％。

可见，并不是“boot”在书面语语料库中出现的比例比在口语语料库中出现的比例大10倍。事实上这个词在这两个语料库中出现的比例相同。因此，在比较两个不同规模语料库的数据时，不应直接比较它们在语料库中的出现次数，而是需要把它们的出现次数进行归一化处理，使它们的数据具有可比性。这时采用的最基本的比例计算公式为：

$$\text{比例} = \frac{\text{词形在语料库中的出现次数}}{\text{语料库中所有词形的出现次数}} \times 100\%$$

这个比例常常用百分数来表示。

三、统计量测试

假设现在要比较马太福音和约翰福音两个拉丁语版本的

圣经，比较的对象是动词"说"(to say)的现在时(dicit)和完成时(dixit)在两种版本中的使用情况。因此需要首先统计这两种动词形式在各自版本中的出现次数。以下是统计的结果：

	dicit	dixit
马太福音	46	107
约翰福音	118	119

从以上数据看出，似乎约翰福音使用现在时(dicit)的次数比马太福音要多。假如这是这两个版本的真正不同的地方，还需要通过统计说明这个结论并非是一个偶然结果，得出这个结论具有多大的可能性。但是单单从上面的次数表无法得到这样的结论，需要进一步的实验，也就是进行统计量测试，以便确定这两个文本在动词"说"(to say)的使用上的不同，究竟有多大可能性是偶然的。

语料库语言学中可以使用多种统计量测试法，这些方法包括 x^2 测试、t 测试等等。为了说明这样的测试在语言分析中的作用，对 x^2 测试做一个简单介绍。x^2 测试是统计量测试中常用的一种方法，具有以下优点：(1)这种测试对数据的敏感性比 t 测试要强；(2)对数据没有"正态分布"的假设，这个假设对一些语言学的数据并不成立；(3)容易计算。x^2 测试的主要缺点是：当被观察的对象在出现次数较少时，测试的结果不可靠。

x^2 测试通常用来比较在语料库中观察到的频次和期望得到的频次之间的差异。期望的频次越靠近观察到的频次，

则观察到的频次就是一个偶然的结果。反之，如果期望的频次和观察到的频次之间的差距越大，那么就越表明观察到的频次确实是受到某些因素的影响而产生的结果，不是偶然的。就上面的例子来说，这两个版本确实在动词“说”(to say)的使用上不同。

略去 x^2 值的计算过程，假设已经计算出它们的 x^2 值，然后在一张 x^2 的数据表中确定 x^2 值的重要性有多大。在此之前需要确定自由度的值，计算公式为：

自由度＝(频次表中列的数目－1)＊(频次表中行的数目－1)

之后，在 x^2 值的表中查找在相应自由度和 x^2 值下对应的概率值。如果这个值接近0，说明统计量并不是偶然结果，反之，可认为是偶然结果。因为概率值在0到1之间，通常设定一个阈值作为统计量有意义还是没有意义的界限，这个值一般定为0.05。如果查表得到的概率值小于0.05，则说明有百分之九十五的把握证明所得的统计量是有意义的；否则所得统计量有意义的可信度不是百分之九十五。

现在再来看刚才例子中的区别在统计数据上是否有意义。用频次表计算出这个表的 x^2 值为14.843。频次表中有两行两列，所以自由度为(2－1)＊(2－1)＝1。通过查 x^2 分布表得到在这样条件下的这个 x^2 值的概率为0.0001，远小于0.05。因此可以判定，这个区别确实反映了这两个版本的圣经在动词“说”(to say)上的使用是不同的，而不是偶然的结果。

四、有意义的搭配

搭配(collocation)是语言学中广泛应用的一个重要概念。简单地说,搭配就是具有一定特征的词语同现模式。Kjellmer (1991)[51]认为:人脑中的词典不仅仅由单个词条组成,其中还包括许多比词更大的语法单元。这些单元中有些是固定的,还有些是可变的。在文本数据中识别词语的同现模式(除了 Kjeller 所说的在语法结构上的模式外,还有与指定词经常同现的那些词),在词典编纂中具有十分重要的作用。它有助于定义一个词的义项以及它们的使用环境。这些信息对自然语言处理和语言教学也同样重要。

从语料库中找出某一给定词的搭配时,既可使用统计的方法,也可以使用信息论中的方法。

1. 互信息和 Z 分值

给定一个文本语料库,可以根据经验数据发现哪些词对之间具有明显的粘合度,这些词对很可能构成有意义的搭配,而并非偶然的邻接。互信息和 Z 分值是经常用来判断一个词对是否可能组成搭配的两种办法。

互信息(mutual information)[52]是信息论中的一个概念。将组成词对(也可以是任意两个被观察对象组成的同现对)的两个词 w_1 和 w_2 的同现,看作一个联合发生的随机事件,互信息的值就是这个联合事件的概率 $P(w_1, w_2)$,除以这两个

词分别出现的概率 $P(w_1)$和 $P(w_2)$，并取对数：

$$M(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

互信息的实际意义是一个词的出现为给另一个词提供的信息量有多大。例如，词对(riding, boot)是一个组合单位；而 formula 和 borrowed，尽管它们也同现过(如句子：It is a formula borrowed from ...)，但这只是一次偶然的同现，它们之间没有特殊的关联。一般来说，两个词目之间的联系越强，它们之间的互信息就越大；如果两个词目是负相关(即一个出现，另一个肯定不出现，反之亦然)，则它们的互信息就是一个负数；如果两个词目相互独立(也就是说它们之间没有关系)，这时互信息为零。换句话说，互信息值较高的两个词目很有可能组成一个有意义的搭配，而互信息值接近于或小于零的词目则不可能构成搭配。

Z 分值(Z-score)提供的数据与互信息数据相似。对文本中某个指定词目来说，Z 分值比较在该指定词目上下文范围内出现的其他词的实际频率与期望频率。一个词的 Z 分值越高，那么这个词与给定词之间的搭配能力(或结合性)就越强，就越可能组成有意义的搭配。在语料库语言学中，Z 分值的方法没有互信息的方法用得多，但是一个名为 TACT 的逐词索引软件就采用了 Z 分值的方法。

2．互信息和 Z 分值的应用

主要用途就是从语料库中抽取多词组合单元，不但包括

像"cock and bull"一样的惯用语，还会含有像"temporal mandibular joint"一样的名词词组。后者抽取的是术语，它除了用于词典编纂外，在专业领域翻译中也有重要用途，可以建立含有一个领域术语的详细知识库。

互信息和Z分值的第二个重要用途就是可以辅助词义辨识(Word Sense Disambiguation，简称WSD)。与前面的应用不同，现在要从语料库的数据中找出一个特定词的最一般的搭配。如果给出某个指定词的一些最重要的搭配，很可能：(1)将相似的搭配组合在一起，从而可以帮助语言学家从大批索引行中自动地识别出该词的不同词义。比如，在英语中"bank"很可能与地理学的词(如river)和金融性质(如investment)的词组成搭配，从而区分出"bank"的两个不同的词义。(2)同时比较具有一定联系的两个不同词的搭配，以确定这两个词在用法上的不同。如里奇(Leech 1992)[53]作过一个实验，比较意义相近的两个词"strong"和"powerful"在用法上的不同。利用互信息从语料中分别抽取出这两个词的搭配，结果发现这两个词的搭配词完全不同。"strong"的搭配词为"northerly"、"showings"、"behaviour"、"currents"、"supporter"等，而"powerful"的搭配词为"tool"、"minority"、"neighbor"、"symbol"、"figure"、"weapon"等。虽然其中某些组合不一定称得上搭配，但是可以说明这两个形容词在用法上的重要差别。

互信息还有一个重要用途，就是帮助定义两个对齐的平行双语语料库的关系。假设存在一个双语的平行语料库，并

且已经完成了句子对齐的工作。也就是说,语料库中一种语言中指定的一个句子,在另一种语言中都有它对应的翻译。而后可以计算得到在这些句子中哪些词是可以互译的。

五、语言模型

1. N元模型(N-1阶马尔可夫模型)

设符号串S由L个符号$w_1, w_2, \ldots, w_L$组成,基于符号同现的语言模型认为,S发生的概率为

$$P(S)=P(w_1)P(w_2|w_1)P(w_3|w_2w_1)\cdots P(w_L|w_1\cdots w_{L-1})=\prod_{i=1}^{L}P(w_i|w_1\cdots w_{i-1})$$

对以上公式建立独立性假设,即假设词串S中的每个词w_i的出现只与它前面相邻的$N-1$个词$w_{i-(N-1)}\cdots w_{i-1}$有关,而与前面N-1词之外的词无关。上面的公式可以近似表示为:

$$P(S)\approx\prod_{i=1}^{L}P(w_i|w_1\cdots w_{i-1})$$

$$=P(w_1)P(w_2|w_1)\cdots P(w_{N-1}|w_1\cdots w_{N-2})\cdot\prod_{i=N+1}^{L}P(w_i|w_l\cdots w_{N-1})$$

N元语法模型相当于N-1阶马尔可夫模型。

最常用的N元语法模型分别是N=2和N=3的二元语法模型和三元语法模型。即近似认为任意一个词出现的概率只与在它前面出现的一个或两个词有关。则因此,相应的概

率计算公式进一步简化为：

二元语法模型的计算公式为：

$$P(S)=P(w_1)\prod_{i=2}^{L}P(w_i \mid w_{i-1})$$

三元语法模型的计算公式为：

$$P(S)=P(w_1)P(w_2 \mid w_1)\prod_{i=2}^{L}P(w_i \mid w_{i-2}w_{i-1})$$

2. 隐马尔可夫模型(HMM)

HMM是由转移链连接的多个状态的集合，其中每个转移链上都有两组概率：转移概率(transition probability)给出了执行该转移的概率，输出概率密度函数(Output Probability Density Function, PDF)定义了在执行某个转移的条件下从有限字母表中输出每个符号的概率。如图3-2所示。

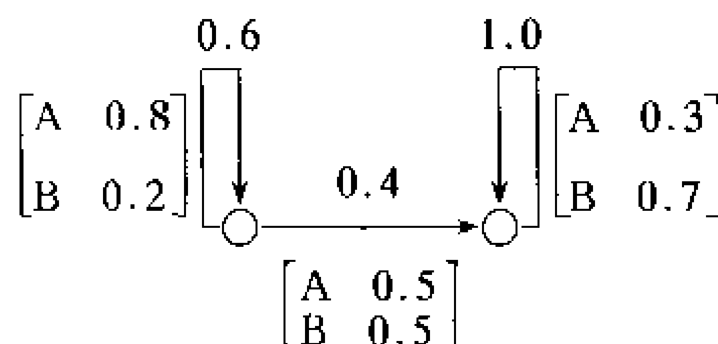


图3-2：一个简单的HMM

(两个状态，两个输出符号A和B)

HMM的形式化定义如下：

(1) 状态集合{S}：包括一个起始状态S_I和一个终止状态S_F；

(2) 转移集合{a_{ij}}：a_{ij}表示从状态i转移到状态j的概

率；

$$a_{ij}=P(X_{t+1}=j \mid X_t=i),\ \forall i,j,a_{ij}\geq 0,\sum_j a_{ij}=1$$

(3) 输出概率矩阵$\{b_{ij}(k)\}$：b_{ij}表示从状态i转移到状态j时输出符号k的概率。

$$b_{ij}=P(Y_t=k \mid X_t=i,X_{t+1}=j),\ \forall i,j,k,b_{ij}\geq 0,\sum_k b_{ij}(k)=1$$

其中，$X_t=j$表示在t时刻处在状态j，$Y_t=k$表示在t时刻的输出符号是k。

给定一个HMM模型M，它产生符号序列y_1^t的概率为：

$$P(Y_1^T=y_1^t)=\sum_{x_1^{T+1}}P(X_1^{T+1}=x_1^{T+1})P(Y_1^T=y_1^T \mid X_1^{T+1}=x_1^{T+1})$$

其直观意义是：列举出所有长度为T且产生y_1^T的状态转移路径x_1^{T+1}(也称为马尔可夫链)，对它们的概率求和，其中每条路径x_1^{T+1}的概率为该路径上转移概率和输出概率的乘积。马尔可夫链X和输出符号序列Y都由一个HMM产生，但是输出序列Y是直接观察到的，而状态序列X是隐含(hidden)的。

在一阶HMM中由两个假设：

①马尔可夫假设(Markov assumption)：

$$P(X_{t+1}=x_{t+1} \mid X_1^t=x_1^t)=P(X_{t+1}=x_{t+1} \mid X_t=x_t)$$

其中X_i^j表示状态序列$X_i,X_{i+1},\cdots,X_j$。马尔可夫假设表示在$t+1$时刻马尔可夫链到达某特定状态的概率只与t时刻

马尔可夫链的状态有关。

②输出无关假设(output-independence assumption):

$$P(Y_t = y_t \mid Y_1^{t-1} = y_1^{t-1}, X_1^{t+1} = x_1^{t+1}) = P(Y_t = y_t \mid X_t = x_t, X_{t+1} = x_{t+1})$$

其中 Y_i^j 表示输出序列 $Y_i, Y_{i+1}, \cdots, Y_j$。输出无关假设说明 t 时刻输出特定符号的概率只与该时刻发生的转移(从 x_t 到 x_{t+1})有关。

在一阶 HMM 模型下,模型 M 产生序列 y_1^T 的概率为:

$$P(Y_1^T = y_1^T) = \sum_{x_1^{T+1}} \prod_{t=1}^{T} P(X_{t+1} = x_{t+1} \mid X_t = x_t) P(Y_t = y_t \mid X_t = x_t, X_{t+1} = x_{t+1})$$

3. N 元语法模型、HMM 和基于语法的模型的比较

N 元语法模型(N－1 阶马尔可夫模型)方法简单,实用性强,被广泛用作语音识别和自然语言领域的语言模型。但是,自然语言最重要的特征是其结构性,而 N 元语法模型是一种基于线性的符号同现关系的语言模型,只能够观察到表示语言最表层信息的符号(一般为字、词或词性标记)之间相邻出现的现象,并不能观察语言的结构,因此用线性的 N 元语法模型来表示结构化的自然语言具有局限性。

HMM 模型是马尔可夫模型的扩展。一般的马尔可夫模型描述的是一个随机过程——状态之间的转移,而 HMM 描述两个随机过程:一个随机过程描述输出符与状态之间的概率关系,即输出符是状态的概率函数;另一个随机过程描述的

才是状态之间的转移关系。作为外界的观察者来说,只能看到输出符,而不能看到状态之间的转移,即状态之间的转移是隐含的。HMM 等价于概率型正规文法[54],是一种有限状态模型,而有限状态不能描述自然语言的层级结构[55]。

4. HMM 模型在语料库中的应用

以语料库的词性标注说明 HMM 在实际语言处理中的应用。

问题描述

假设有一个词序列 $w_1, w_2, ..., w_T$,现要找一个词性序列 $c_1, c_2, ..., c_T$,由于词性歧义现象普遍存在,所以一个词序列可以对应不同的词性序列,而我们现在要找的词性序列是使公式 $PROB(c_1, c_2, ..., c_T | w_1, w_2, ..., w_T)$ 的值最大的词性序列[56]。

利用概率中的 Bayes 公式,上式可写为:

$$\frac{PROB(w_1, w_2, ..., w_T | c_1, c_2, ..., c_T) \times PROB(c_1, c_2, ..., c_T)}{P(w_1, w_2, ..., w_T)}$$

其中,$PROB(w_1, w_2, ..., w_T | c_1, c_2, ..., c_T)$ 称为词汇知识公式,$PROB(c_1, c_2, ..., c_T)$ 称为语言模型,而对同一词串,分母的值都相同,所以上面的公式可以简化为求使下列公式值最大的词性串:

$$PROB(w_1, w_2, ..., w_T | c_1, c_2, ..., c_T) \times PROB(c_1, c_2, ..., c_T)$$

可以对这个公式作进一步的假设:当前词的概率只取决

于该词的词性，当前词的词性只与以前词的词性有关。最终整个问题就可以利用下面公式表示为：

$$T^* = \underset{c_1,c_2,\ldots,c_T}{\operatorname{argmax}}\, p(w_1)\,p(w_1 \mid c_1)\,p(c_1)\prod_{i=2}^{T} p(c_i \mid c_1, c_2, \ldots, c_{i-1})\,p(w_i \mid c_i)$$

T^* 为最终标注的词性序列，$p(.)$表示概率。

从以上公式就能够推出一阶或二阶的 HMM。

一阶 HMM 是设当前词的词性只与它的前一个词的词性有关。具体公式为：

$$T^* = \underset{c_1,c_2,\ldots,c_T}{\operatorname{argmax}}\, p(c_1)\,p(w_1 \mid c_1)\prod_{i=2}^{T} p(c_i \mid c_{i-1})\,p(w_i \mid c_i)$$

其中 $p(c_1 \mid c_{i+1})$为 HMM 中的状态转移概率，$p(w_i \mid c_i)$为词汇生成概率。

这样词性标注问题就表示成为一个在一阶 HMM 上求最佳路径的问题。对上面状态转移概率和词汇生成概率可以通过对已标注了词性的语料库训练得到。

第三节　逐词索引软件及其应用

本节介绍由 Sinclair 开发研制的两个用于语料库开发的统计软件，这两个软件都是逐词索引软件，为指定词提供它在语料库中每次出现时的上下文环境。不同的是，这两种索引软件将检索到的索引行用不同的统计方法将统计数据进行排

序,帮助人们对语言现象进行分析。Collocate 计算与指定词同现的那些词的相对明显性,Typical 则计算整个检索行的明显性。这两个程序的使用效果都超出了最初的期望,在语料库开发中具有普遍的指导意义。文献[57]详细介绍了他在这两个软件上所做的工作,下面给以介绍。

一、COLLOCATE 程序

Collocate程序首先计算一个词在逐词索引文件中的频次,以及它在语料库中的出现频率,然后计算这个词与指定词构成搭配的明显性。这里明显性的定义是候选搭配词在整个语料库中出现的概率与它在逐词索引文件中的出现概率之间的比值。

1. 算法

对给定的词,首先找出这个词在语料库中的每次出现,然后对构成的逐词索引文件中每个索引行中的每个词,计算它的观察频率和期望频率。在计算时,可以选择如下不同方法来对他们进行记频:不计英语单词的大小写,可以考察去掉屈折变化后的词型,也可以只考察指定词的左边、右边或左右两边的搭配等。

算法的输入:设逐词索引文件为 concordance file。当指定词在语料库中出现时,由出现在该词左右一定长度的窗口内的上下文组成一个索引行。所有索引行组成了这个指定词

的逐词索引文件。

首先从大规模语料库中统计得到全部词条的频率表。

对每个与指定词同现的词 w，计算它与指定词构成搭配的明显性 S：

$$S=\frac{w\text{ 在索引行文件中的出现概率}}{w\text{ 在整个语料库中的出现概率}}=\frac{OF}{EF}$$

其中，$OF=freq_{span}(w)/N_{span}$，$EF=freq_{corpus}(w)/N_{corpus}$，$freq_{span}(w)$和 $freq_{corpus}(w)$分别表示词 w 在索引行文件和整个语料库中的出现次数，N_{span}和 N_{corpus}分别为索引行文件和语料库中所含的词次数。

输出：输出时将与指定词同现的所有词按照其明显性的值排列，每行包含四列信息，它们分别为：

搭配词；与指定词同现的搭配词。

词频；该搭配词在语料库中的出现次数。

期望频率；在指定上下文长度内，这个搭配词可能出现的期望频率。

真正频率；搭配词在逐词索引文件中的实际出现次数。

执行这个程序时，可以有两个选项：(1)是否区分大小写；(2)是否包含位置信息。以下用三个具体的实例来说明这个程序的应用。

设现在考虑的词为“arms”，上下文长度为前后 4 个词，表 3－1 和表 3－2 分别给出了不计词的屈折变化和考虑屈折变化两种情况的输出结果。其中表中第一列为与指定词同现的

搭配词，第二列和第四列分别是搭配词在语料库中和在索引行文件中的出现次数。第三列是搭配词经过程序计算后得到的明显性的值。以后凡使用程序 Collcate 后给出的结果表各列的意义同上。

表 3-1 不计词的屈折变化

Caches	75	0.073	30
Outstreched	372	0.364	97
Cache	254	0.248	66
Cradled	168	0.164	38
Flailing	237	0.232	37
Embargo	3427	3.352	527
Folded	1462	1.430	195
Ammunition	1910	1.868	154
Shipments	1238	1.211	96
Treaties	807	0.789	60
Legs	8507	8.320	594
Waving	1741	1.703	116
Aloft	409	0.400	23
Torso	406	0.397	21
Flung	1176	1.150	59
Elbows	613	0.600	30

表 3-2 考虑词的屈折变化

Buildup	441	0.431	21
Cache	329	0.322	96
Outstretched	372	0.364	97
Flail	3791	3.708	556
Ammunition	1910	1.868	154
Aloft	409	0.400	23

(续表)

Fold	4707	4.604	236
Cradle	1085	1.061	53
Buildup	441	0.431	21
Torso	475	0.465	22
Strategic	5712	5.587	256
Fling	1857	1.816	83
Reduction	8191	8.011	343
Conventional	7157	7.000	292
Gent	615	0.601	25
Smuggle	2024	1.980	78

从表 3－2 看出,"cache"是与指定词"arms"的明显性最高的一个词,它在语料库中的出现次数为 329,从表 3－1 中发现,它实际上是"cache"和"caches"这两个词形在语料库中出现次数的总和。表 3－2 中的搭配词是把同现词去掉屈折变化后得到的词型。

表 3－3 是考虑搭配词位置后的搭配信息表,这时需要增加一列,即第五列,表明这个搭配词通常出现在指定词的哪个方向上。

表 3－3 考虑词的搭配方向(左搭配和右搭配)

Caches	75	0.073	27	No left
Cradled	168	0.164	36	No right
Outstretched	372	0.364	65	Left discarded
Embargo	3427	3.352	488	Left discarded
Embargo	254	0.248	66	
Ammunition	1910	1.868	153	No left

(续表)

Treaties	807	0.789	57s	No left
Folded	1462	1.430	195	
Shipments	1238	1.211	82	No left
Waving	1741	1.703	105	No right
Aloft	409	0.400	23	No left
Legs	8507	8.320	468	Left discarded
Strategic	5712	5.587	249	No right
Reductions	1888	1.847	82	No left
Lifting	2880	2.817	122	No right
Conventional	7157	7.000	276	No right
Reduction	6303	6.165	241	No left
Supplying	1311	1.282	50	No right
Flung	1176	1.150	44	No right
Negotiator	913	0.911	33	No left
Explosives	1257	1.229	40	No left
Shipment	7371	0.721	22	No left

表中,"no left"和"no right"分别表示该搭配没有在指定词的左边或右边出现,如果一个搭配词在一边的出现次数是另一边出现次数的三分之二以上,在另一边的出现就被放弃。"left discarded"和"right discarded"就表示左边或右边被放弃。如果搭配词在两边出现的次数相差不多,则该列的值为空。

二、TYPICAL 程序

Typical 程序在计算索引行内所有同现词的明显性的基

础上，估计整个索引行的明显性，这样有助于找出有特性的实例行。程序的输入是从语料库中得到的逐词索引文件，以及在这个语料库上的一个词汇频率表。之后程序处理这个文件中的索引行，找出其中最典型的索引行。这个程序最初的设计意图是想找出一个指定词的一些典型引用实例。目的是为词典编纂者提供一个有用的辅助工具，以便找出这个指定词的可靠有用的实例。但是，在这个程序的实际应用过程中，意外地发现这个索引程序对词义排歧也很有帮助。

这个程序在设计时引用的一个假设是：文本中的每个词都在一定程度上与其他词相互吸引。这个程序就是要在一个逐词索引文件中找出被这个指定词吸引的全部搭配词。

1．算法

输入：(1)大规模语料库中的词频表；

(2)从这个语料库中获取的指定词的逐词索引文件。

输出：按照典型度值的高低对逐词索引文件中的索引行进行排序的结果。

过程：对索引行的上下文中的每个词 w_i，如果它的出现次数超过某一阈值，则计算：

$$x_i = \frac{P_s}{P_c} = \frac{freq_{span}(w_i)}{N_{span}} / \frac{freq_{corpus}(w_i)}{N_{corpus}},$$

其中，p_s 为 w_i 在索引行文件中一定上下文长度内的出现的相对频率，p_c 为 w_i 在整个语料库中出现相对频率，即大规模

语料库中词频表中的频率。

然后，用 *Z-score*（z 分值）进行归一化：

$$\bar{x}=\sum_{i=1}^{n}\frac{1}{n}x_i,$$

其中，n 为大于某一阈值的与指定词同现的并且同现次数大于一定域值的所有词的个数。计算每个同现词的 z 分值：

$$z=\frac{x_1-\bar{x}}{s},$$

s 为标准偏差，计算公式如下：

$$s=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

最后，将每个索引行中的与指定词同现的所有词的 z 分值相加，就得到有关这个索引行一个数值，称为这个索引行的典型度。将这些索引行按典型度由大到小的顺序排列输出。输出的格式是索引行以及该索引行对应的典型度值。具有相等典型度的索引行进一步按字母顺序排列。

这里需要说明：一个索引行的典型度是这个索引行中的每个同现词的 Z 分值的累加和，反映出组成这个索引行的各个成分的权值。如果一个索引行有一个词的 z 分值较大，则这个索引行的典型度就较高，并且具有同一个较高 z 分值的词的不同索引行应该具有近似的典型度。但是利用标准偏差计算典型度的缺点是：如果有些词的出现频率与平均的出现次数相等时，这些词的 z 分值为 0。为处理这种情况，在找到具有典型度高的索引行后，辛克莱采用了从逐词索引文件中

去掉排在前面的高典型度的索引行后，再重新调用上面算法找出另外的具有较高典型度的索引行。

2．实例

辛克莱等人利用这个程序对某些词的搭配进行了研究，设选定的词形为“hot”，这个词形有多种意思，可以表示天气很热，或是一种古怪的气味等。

Typical 程序将“hot”的逐词索引文件中的索引行按照每个索引行的典型度排列，希望排列在一起的索引行中出现相同的搭配词，特别是那些具有很高的 z 分值的搭配词所在的索引行应当排列在一起。这个程序选定的索引行的上下文长度为前后 3 个词，搭配词在索引文件中的最少出现次数为 7，“hot”的逐词索引文件及频率表是在含有两亿词次的英语语料库上完成的。图 3－3 给出了排列结果。图中每行前一项是索引行对应的典型度，“<>”之间是这个索引行。

19476.18<a sackful of guitars shaped like red hot pokers that stab the songs through>

19476.18<a sackful of guitars shaped like red hot pokers! Visual fireworks: STEVE>

19476.18<palms and pampas grass among the red hot pokers seem like a feasible and>

19474.55<like lupins and delphiniums, red hot pokers (Kniphofias), mullein >

19474.55<had gardens with raspberries and red hot pokers. Once we spent a week in the>

19474.55<their weed-choked snapdragons and red hot pokers. If they ask about it, smile>

…………

15446.13<doctors found it helped dry vagina, hot flushes, sweats, ftension, anxiety and>

15082.65<Problems of the menopause such as hot flushes , night sweats, dry vagina>

15076.28<< FCH > symptoms, which include hot flushes, sweats, tingling, and >

15070.74<in particular can help with hot flushes, night sweats, vaginal >

…………

13807 29<an urgent need for the bathroom, hot and cold flushes and pins and >

13790.39 turn white and that know what I mean. Hot and cold flushes and that. <M01>Was>

13790.39 and pins and needles < FCH > < M38 > < FCH > hot and cold flushes, sweating, >

13579.00<service was held on a blisteringly hot day. When the body was lifted from>

13574.27<< LTH > Sunday was blisteringly hot with cars and drivers alone >

13539.29<designed to combat such blisteringly hot days whereas Malcolm roars like an>

13517.17<said : We have enjoyed a blisteringly hot June, along with Scandinavia, the>

…………

9875.08 <in. This is especially a danger in hot and humid climates, such as the >

9866.32 <known as'dropouts', when played in hot and humid climates. <LTH> Examination>

9676.71 <Phil found himself slaving over a hot grill at minimum wage while being>

9027.15 <Summers can be unbearably hot and humid and the scenery is flat >

8424.17 <<CQ1>I don't like slaving over a hot stove cooking a good meal and>

8341.5 we spent hours lovingly slaving over a hot stove making, it's hardly surprising>

8334.75 <but when you are slaving over a hot stove in the kitchens of the Hotel>

8334.75 <<t>WOMEN spend hours slaving over a hot stove in the kitchen but are banned>

…………

8135.44 <driving the criminals' favourite hot hatchbacks cars # who face increases>

7822.09 <rising insurance costs, even more 'hot hatchbacks' like the nwe Citroen ZX>

7820.46 <among the new Classics are the hot hatchbacks and homologation specials>

7820.46 <will go straight into group 20, and hot hatchbacks can expect to see their>

7820.46 <borne by those owning what they call hot hatchbacks and sports cars. Er the>

7816.97 <on 45 high risk models, mostly hot hatchbacks, was swift. New Escort RS>

…………

7356.70 <display, and where summers can be hot anc humid, sow in September where>

7120.72 sheets. (The dough is soft, so in very hot or humid weather, refrigerate it for>

7074.93 <tin. <t>Baking Note: <FCH> In very hot or humid weather, or if your kitchen>

7037.53 <a day for four days in weather so hot and humid that several men died. He>

7035.44 <and maintain a humid stmosphere in hot weather. Keep it shaded form the>

7020.37 <is partly affected by the weather. On hot humid days large amounts of pollen>

7005.79 <attack even without exercise. Very hot or humid weather will make it>

…………

5465.49 <down because you stay dry and have a hot toddy when you get home>

5462.31 <a sherry at the theatre bar, or a ' hot toddy ' to help keep the cold at bay. >

5448.56 <He dipped his tiny beak into a hot toddy after this series of knight's>

5448.26 <you head off down the slopes - it ' s a hot toddy with an extremely potent kick>

5440.70 <but the Club were as cheering as a hot toddy. Even though they seemed>

5414.71 <leader Paddy Ashdown, branded the 'hot toddy' budget as a cynical manoeuvre>

…………

图 3 3 根据行的典型性计算对"hot"的索引结果

从图 3-3 中看到"hot"的常用搭配"hot flushes"及"hot and humid"等。从图中还可以看出索引行中典型度的每次在数值上的明显变化都象征着"hot"这个词的一个新的用法。因此,索引行的典型度的变化可以表征指定词的一种新用法的开始。

3. 程序中几个变量对语言分析的影响

在上面两个统计程序之前,首先需要用户输入的几个变量的取值。首先是指定使用本次执行哪一个统计程序(Collocate, 或 Typical)。之后由用户指定所研究的具体词,以及参与计算的索引行文件中索引行的数目。不同规模的文件产生不同的结果。此外,在输入时还需要说明,在建立索引行时指定词左右两边的上下文的长度,以及搭配词在这个上下文范围内出现的最少次数(阈值)。下面是辛克莱在运行这两个程序时得到的一些经验。

(1)输入文件的大小

两个统计程序对逐词索引文件的大小没有限制，文件的规模越大，得到的统计数据就越可靠。表 3－4，表 3－5 和表 3－6 给出索引文件分别为 5000 行、20000 行和 50000 行时运行 Collocate 程序产生的搭配统计结果。这时程序选定的上下文长度都是前后 4 个词，搭配词的最少出现次数为 10。

表 3－4 索引行个数为 5000 行时的运行结果

Contorted	161	0.039	11
Flushed	691	0.167	11
Grin	1329	0.321	19
Mask	2221	0.536	26
Fines	1260	0.304	11
Starvation	1421	0.343	12
Smile	7371	1.780	54
Staring	2677	0.646	19
Brave	3549	0.857	22
Pale	4456	1.076	27
Expression	5760	1.391	31
Smiling	3123	0.754	14
Handsome	3263	0.788	13
Charges	15155	3.660	60
Face	49108	11.860	185
Tears	5429	1.311	19
Buried	4079	0.985	13
Neck	7236	1.747	22
Touched	3993	0.964	11
Prospect	6612	1.597	17
Value	21732	5.248	56
Thin	7341	1.773	17

表 3－5　索引行个数为 20000 行时的运行结果

Volte	69	0.067	23
Expressionless	165	0.159	29
Impassive	171	0.165	24
Contorted	161	0.156	22
Sallow	117	0.113	12
Ashen	120	0.116	12
Adversity	392	0.379	34
Creased	163	0.157	14
Flushed	691	0.668	51
Craggy	167	0.161	11
Frown	415	0.401	24
Haggard	260	0.251	15
Shadowed	234	0.226	12
Slap	910	0.879	46
Tanned	499	0.482	24
Slapped	764	0.738	36
Bony	308	0.298	14
Brightened	282	0.272	12
Streaked	284	0.274	12
Ruddy	285	0.275	11
Screwed	634	0.612	23
Beaming	420	0.406	15

表 3－6　索引行个数为 50000 行时的运行结果

Volte	69	0.167	53
Barroom	34	0.082	18
Expressionless	165	0.398	73
Eiger	30	0.072	13
Contorted	161	0.389	60

(续表)

Impassive	171	0.413	56
Broderick	89	0.215	24
Ashen	120	0.290	28
Puckered	89	0.215	20
Reddened	111	0.268	25
Blotchy	59	0.142	13
Freckled	104	0.251	22
Creased	163	0.394	32
Redder	64	0.155	12
Puffy	161	0.389	29
Adversity	392	0.947	66
Flushed	691	1.669	112
Sallow	117	0.283	19
Slap	910	2.198	149
Sunburned	69	0.167	11
Smirk	170	0.411	23
Craggy	167	0.403	22

由这三个表看出，用规模较大的逐词索引文件得到的搭配更具有代表性。

(2)上下文的长度

这是另一个需要用户说明的变量。在英语中人们常常把上下文长度限定为指定词的前后4个词。一个词对它出现的上下文具有一定的影响，也可以说，一个词的上下文是这个词的意义的一个标识。如果一个词是多义词，那么这个词的上下文就可以表明这个词此时的词义。因此，有必要确定一个词所影响的上下文一般是前后多少个词，以便选择能产生最

佳分析结果的上下文长度。

如果对一个指定词运行 Collocate 程序，每次只改变上下文的长度，而保持其他条件不变，则每次将得到的输出结果是不同的。表 3-7 和表 3-8 均以"eye"为指定词，逐词索引文件的大小均为 5000 行，并且搭配词的最少出现次数为 10 次，但上下文长度分别为前后两个词和前后 6 个词。以下是这两种上下文长度的运行结果。

表 3-7 只取上下文中前后两个词的运行结果

Beady	99	0.012	12
Watchful	346	0.042	35
Remover	174	0.021	14
Untrained	249	0.030	13
Sockets	296	0.036	12
Socket	464	0.056	12
Blind	4941	0.097	81
Catches	1607	0.194	21
Naked	3486	0.421	40
Gel	1054	0.127	12
Caught	14201	1.715	133
Eagle	1912	0.231	18
Witnesses	3304	0.399	25
Eye	16359	1.975	121
Keeping	12294	1.485	76
Meets	5019	0.606	29
Contact	16184	1.954	90
Patch	2173	0.262	11
Witness	4342	0.524	21

（续表）

Catching	2510	0.303	12
Keep	48681	5.878	211
Catch	8595	1.038	36

表 3－8 取上下文中前后 6 个词的运行结果

Beholder	119	0.043	24
Beady	99	0.036	13
Remover	174	0.063	20
Watchful	346	0.125	36
Glint	241	0.087	21
Twinkle	249	0.090	18
Contour	195	0.071	13
Retina	221	0.080	14
Untrained	249	0.090	13
Glam	295	0.107	14
Sockets	296	0.107	12
Blink	397	0.144	15
Socket	464	0.168	16
Gel	1054	0.382	20
Blind	4941	1.790	90
Catches	1607	0.582	26
Eye	16359	5.926	242
Makeup	826	0.299	12
Naked	3486	1.263	43
Caught	14201	5.144	154
Keeping	12294	4.454	18
Eagle	1912	0.693	18

从表中看出，搭配词“beady”、“watchful”和“remover”等在两个表中都具有很高的明显性，然而还发现，表 3－7 中一

些具有较高明显性的搭配词（如"patch"和"witness"），在表3－8中消失了，或只具有较小的明显性。这是因为，在长度为2的上下文中具有较高明显性的搭配词，也一定在长度为6的上下文中出现。因此。它们的明显性将被在更大上下文中出现的其他词所削弱。而只有那些明显性很高的搭配，或那些出现位置相对自由的搭配，才被保留在上下文长度为6的搭配表中。像"patch"只可能出现在"eye"的后一个词的位置上，因此它的明显性就被那些在长度为6的上下文中常出现的搭配减弱了。在表3－8中，"beholder"具有最高的明显性，而这个词却没有在表3－7中出现，这是因为"beholder"常在短语"in the eye of the beholder"中出现。这时"beholder"在指定词"eye"的前后两个词之外。为避免这种情况，可以在运行程序时加入位置信息，比如只考虑指定词前面的词或后面的词，这时搭配的明显性就会提高。

(3)搭配词的最少出现次数

搭配词的最少出现次数是确定参与计算的搭配词在逐词索引文件中出现的最少次数。该变量直接控制进入搭配词的候选个数。如果这个阈值较小，程序就需要较长的运行时间，并且可能产生某些错误的结果；反之，如果这个阈值定得较大，一些具有较高明显性的搭配就可能被忽略。

设定搭配词的最少出现次数，是为保证程序运行时不考虑那些在语料库只出现过一两次及在语料库中发生拼写错误的词、人名或专有名词等。从表3－9中就可以看出，设定搭

配词的最少出现次数的重要性。表中找到的具有最高明显性的词在语料库中只出现过两次，而在“hard”的上下文中就出现了一次，因此它们具有高明显性。从表中看出，“getthem”是一个输入错误，中间少了一个空格。正确的输入应该为“get them”。

表 3－9 设定最少出现次数对程序运行结果的影响

Anic	2	0.000	1
Bogna	2	0.000	1
Endochorion	2	0.000	1
Getthem	2	0.000	1
Givada	2	0.000	1
Hipp	2	0.000	1
Kinjiro	2	0.000	1
Korbel	2	0.000	1
Leinoff	2	0.000	1
Lektropaks	2	0.000	1
Leshenka	2	0.000	1
Maternite	2	0.000	1
Mogulled	2	0.000	1
Pittesburg	2	0.000	1
Pogrebnjak	2	0.000	1
Sarit	2	0.000	1
Shirtlifter	2	0.000	1
Spener	2	0.000	1
Tolars	2	0.000	1
Trancepotter	2	0.000	1
Weasling	2	0.000	1

一般来说，这个变量的取值应该随着上下文长度的变化

而变化。当上下文长度取得较短时,这个值就较小;随着上下文长度的增加,这个值应逐渐增大。

第四节 语料库标注

收集一种语言的大量文本并储存在计算机上,形成一个大规模的语料库。之后,研究者希望从这个语料库中抽取所需的信息。例如,利用真实的语言数据来建立一种语言更好的词典或语法手册,以便成功地理解和使用这种语言。为了从不同语料库中抽取信息,必须首先在一个或更多层面上对语料库进行分析,并且将分析结果标注到语料上去,从而给一个语料库带来附加的价值。这就是语料库标注。语料库标注作为对语料库的关键贡献,已被广泛接受。文献[58]详细论述了对语料库进行各个语言层面的标注方法,在我们这套丛书中也有有关汉语语料库标注的方法的叙述,这一部分并不涉及语料库标注的算法,只介绍语料库标注的含义、标注规范和语料库标注的范畴。

一、什么是语料库标注

语料库标注可定义为:是一种给口语和(或)书面语语料库增添解释的(interpretative)和语言的(linguistic)信息的实

践。“标注”也可以指这个过程的最终产品，即附加或分散在语料库中的语言标记。语言标记可以是音节的节律标记、词性标记、句法标记、语义标记等。由于汉语的词之间缺少空格，所以还有词的标记，一般用一个空格表示。语料库标注的一个最典型和最著名的、也是在语料库研究最赋有影响力的例子就是语法标注，也称为词类标注、词性标注或 POS 标注。在这种标注过程中，对语料库中的每个词(次)附加一个标记或符号，以指明这个词的语法类，也就是词类。如“公布/vgn”中，语法标记“vgn”表示“公布”在这里是一个带名词宾语的动词。

把标注称为是“解释的”信息的原因，是因为标注至少在某种程度上，是人们对文本作出理解的产品。对汉语文本进行分词，表明已经从一个连续的字串中分别出词，进行词性标注，则表明每个词在文本中所具有的词性信息等等。其次，在文本的“标注”和“表示”之间是有区别的。可以通过一个书面语文本来区分这两类信息。一个文本的正词法记录是一个罗马符号的序列，其中带有标点和空格等。这种记录在计算机中可以用特定的代码来进行电子化的表示，在这些代码和可视符号之间存在着一对一的映射，即原始的正词法记录被一个电子文档来表示。在这个表示过程中可能丢失某些细节的信息，比如字体和字号。但这是允许的，因为这类信息对于文本作为一种语言现象的表示来说不是本质的。与此相反，一个文本的标注是元语言的(metalinguistic)：它提供的是关于

该文本的语言信息,而不是文本本身的内容。

但是对于一个口语话语来说,有时很难区分表示的和解释的信息。在把口语转写成书面语或电子形式的过程中,一个转写者在表示该话语的过程中,必然伴随有某种解释。大多数转写出于方便只对例外的发音采用语音描写来辅助传统拼写的词,但这仅仅给出了语音事件的表层可读性,它的物理、语言或社会正式性质也许复杂和难琢磨得多。例如重音和音调节律的标注在一定程度上取决于转写者的判断与经验,同时也取决于所采用的分析系统。

二、为什么要标注语料?

1. 信息获取

只有当语言研究者能够从语料库中获取知识或信息时,才能说这个语料库是有用的。事实上,为了从语料库中抽取语言信息,必须首先向该语料库中植入信息—即添加标注。没有经过任何处理的电子文本语料库称为生语料库(raw corpus)。这样的语料库尤其是汉语的生语料库包含很少关于词法和语法等的信息,因此其应用价值就很有限。例如:在英语中,词 left 作为一个形容词,是 right 的一个对义词,如"my left hand";同时也可以作副词:"turn left";或名词:"on your left"。但作为 leave 的过去式或过去分词,它是一个动词,如"I left early"。因此,left 是一个有多种用法的词。但

是它的各种意思和用法在不带任何标注的生语料中却不能明显地被标示出来。这样的语料库作为词典编纂的资源，是有缺点的。如果语料库经过成功地语法标注，left 的每次出现都有一个标记来表明它的词性，这些信息可以帮助改进词典的编纂。又如在文本－语音转换（Text To Speech）的应用中，英语词 lead 作名词时发音是/led/，作动词时发音是/li:d/。如果要研制一个语音合成器（把输入文本转换成语音输出），那么这个合成器有必要区分 lead 是名词还是动词，以便产生正确的发音。而在汉语中多音字现象也普遍存在，如："行"在"银行"中发音为"hang2"，而在"行人"中的发音为"xing2"。这时在进行文本—语音转换时，也需要为"行"标上正确的读音。经过语料库的词法和语法等信息的标注将给这种语音合成器提供所需的信息。

2. 复用性

语料库的复用性指带标注的语料库资源可以被多次重复使用。可能会有人认为，不需要对语料库做穷尽的标注，而只需要设计一个能够辨识词性的程序。比如 left 在名词前是形容词，在动词后是副词等。但这样做有两个缺点：

（1）根据以上例子，为了辨识目标词的词性，需要事先知道它邻近词的词性。因此，词类的辨识实际上不能看作是一个孤立的问题。

（2）进行语法标注和其他层次上标注的宗旨是：一旦把标注加到语料库上，所产生的新语料库便是一种比原语料库

更有价值的资源,并且还可以提供给其他用户使用。语料库的标注一般是昂贵和耗时的,但如果能复用就物有所值了。

3. 功能多样化

一个经过标注的语料库通常有许多不同的目的或应用,这就是这样的语料库的多功能性。前面已经提到语法标注有两个不同的应用—词典编纂和语音合成。它还可以在语言工程的其他方面得到应用,如机器辅助翻译或信息检索等。因此,标注给语料库增添了一般意义的"附加价值"。而语法标注作为一种最基本的标注,是向更高层次的标注,如句法标注和语义标注,迈出的第一步。由于存在着许多不同目的的用户利用那些带标注的语料,其中有些目的甚至是原标注者未曾想到过的,这使得带标注语料库的复用性显得更加重要。

三、语料库标注的规范

语料库标注者的"专家"水平,以及他们所采用的标注规范的合理性和可用性,决定了语料库标注的信息是否有用,是否有知识。在语料库标注的短暂历史上,对语料库建造者所加的标注,其他人用起来很困难,或甚至不能用的例子时有发生。为了避免这种情况,应该对语料库标注课题实施以下的指导原则:

1. 应当总能在去掉标注后恢复到原来的生语料库。也就是说,生语料库是可以复原的。

2. 标注信息应该能从语料库中单独抽出，有可能的话可以独立存储。

3. 语料库的用户能够方便地查阅到包含下列信息的文档：

(1) 标注规范—即标注所用标准的描述和解释性文档。

(2) 记录标注者、标注地点和怎样标注的文档。

(3) 由于标注通常会出现差错、不一致或歧义现象，因此应当有关于标注质量的说明。例如，语料库的标注结果被校核到什么程度，它的精确率有多高(被判断为正确标注的百分比)，以及标注的一致性达到什么程度等。

4. 应当向用户声明，语料库的标注并非绝对无误，它只是一种在一定意义上有用的资源。标注规范也仅仅是为了在实践中提供一种有用的依据。许多用户会发现使用已经标注好的语料库是值得的，这比用他们自己设计的规范去实现标注好得多。因为这是一项需要花费许多年才能完成的任务。

5. 为了避免误解和误用，标注规范最好尽可能地建立在大众公认和理论上中性的数据分析基础上。尽管标注肯定会面临某些理论上敏感的决策，但它们的目标是应当尽量被广泛接受和理解。

6. 任何标注规范都不能作为“第一标准”。在实践中标注规范倾向于变化。例如，被标注的语料库的规模可能会妨碍过分详细的标注。标注的主要目标需要优先考虑某些类型

的信息等。

尽管有如上6项原则，仍然有些人主张在语料库标注中制定某种标准化，而且似乎在今后几年的实践中会逐步达到某种程度的统一。标准化的原因之一是惯性(inertia)：如果人们发现某种标注规范有用，就会坚持用这种标注规范来开发自己的带标语料库。另一个原因是已经强调过的复用性原则。如果不同研究者需要交换数据和资源(如带标语料库)，不同单位用相同的标准或指导原则，上述交换就容易实现。在需要相互交换语料库软件工具时，标注的某种标准化的需要就更明显了。

四、语料库标注的范围

汉语的语料库标注范围与英语的语料库标注范围略有不同。原因是，在汉语的文本中，词与词之间没有空格标示。把文本中的字串分隔成词串，就是汉语分词标注的任务。汉语的分词是汉语自然语言处理中一项重要的基础工程，是实现以词作为文本处理单位的整个信息处理过程的一个必经阶段。汉语信息处理包括了以词为索引单位的信息检索、信息抽取、机器翻译和句法分析等。经过近20年的努力，当前汉语自动分词技术已经取得了很大的成绩，分词的正确率已可达到99%左右[59]。但是，在汉语文本的分词中还有一些问题没有得到根本解决。它们包括：人名、地名、机构名等专名

和未登录词的识别，以及交集和组合歧义字段的切分等。

以下各部分分别介绍各种语言中普遍存在的语料库标注过程：语法标注（或词性标注）（grammatical tagging）、句法标注（syntactic annotation）、语义标注（semantic annotation）和话语标注（discourse annotation）。

1．语法标注

第一个语料库标注的项目是1971年在布朗大学的布朗语料库上进行的。在两名语料库语言学家Francis和Kucera的指导下，由两名硕士研究生采用上下文有关规则的方法完成。其语法标注集中含有77个语法标记。这些语法标记不仅能够识别词性，如名词、动词和形容词等，还能给出他们的子类，如名词的单复数，形容词的一般级、比较级和最高级等等。

这个语法标注程序的标注正确率达到77%，由人工改正标注中的错误，最后得到了一个非常有用的产品——带语法标注的布朗语料库。这项研究工作的意义在于它第一次展现了语料库标注的一般特性。即一方面显示了自动标注和人工标注之间的区别：采用自动标注方法的必要性，但随后又必然伴随大量的人工核校工作。人工标注和自动标注过程是一项渐进的循环过程，纯粹的人工标注是不可行的。另一方面，只有当自动标注具有足够高的标注正确率时，它才是可用的。

第二个语法标注课题是1982年在LOB语料库上完成的。与前面不同的是：它采用了基于语料库的概率统计方法。

它利用带语法标注的布朗语料库作为统计资源，来计算任意两个语法标记在 LOB 语料库中的转移概率等参数。这个被称为 CLAWL 的语法标注器的标注正确率达到 96.7%。也就是说，和基于规则的第一个语法标注器相比，标注正确率一下子提高了近 20 个百分点。

在这之后，陆续涌现出许多语法标注器，大部分都采用了概率方法。这种方法的主要问题是需要有一个带标记的训练语料库，并且在概率计算时上下文的长度受限，即一般只考虑了指定词左边(或右边)的一个或两个词。

汉语的语法标注研究开始于 20 世纪 80 年代末和 90 年代初，最早进行汉语语法标注的研究单位是清华大学和山西大学。

2．句法标注

句法标注就是给文本中的句子加上它们的句法结构信息。这个思想最初是由 Ellegard[60]提出的，即研制带句法标记的语料库。他和他的学生在 1978 年一起以手工方式对布朗语料库的一部分(含有 128000 个词次)进行了句法分析。到 80 年代，奈美根(Nijmegen)[61]和兰开斯特(Lancaster)[62]大学分别开始建立对语料库实施句法分析的系统。在 90 年代初期，树库(Tree Banks)的建立表明，带句法标注的语料库是自然语言处理研究中的一个重要资源。比如，在语音识别和机器辅助翻译中，都需要有鲁棒的句法分析器。

Lancaster/IBM 的树库规模为三百万词次，宾西法尼亚

大学树库[63]的建立又为这类资源带来更广泛的用户。John Hopskin 大学的 Chellba 利用这个树库建立的基于句法结构的语言模型[64]，有效地解决了语言模型中远距离搭配问题，初步实验取得了较高的识别正确率。树库这个术语表明，短语结构树是语料库句法标注的基本模式。句法标注是一项比语法标注更复杂、对资源要求更高的工程。因此它的研究在许多方面滞后于语法标注，如开始的时间晚，不很成功，并且标注正确率不高等。

对语料库进行句法标注时，对句子可以进行完全的或部分的句法分析。经过句法标注的语料库有以下这些潜在的用途：

(1) 研制句法分析器

进行句法标注的一个重要用途就是用来训练和测试句法分析器，而句法分析器又是许多自然语言处理应用中的一个不可或缺的关键部件。使用带句法标注的语料库有助于建立概率型句法分析器，增加句法分析器的鲁棒性。Jelinek 和 Collins 在宾州大学的树库上建立了一个概率的句法分析器，具体方法见文献[65]和文献[66]。

(2) 抽取词汇信息

带有句法标注的语料库含有丰富的词汇－语法信息，有助于建立计算机词典。计算机词典是一种有结构的资源，为自然语言处理提供所需要的有关词的形态变化、句法和语义等信息。使用带句法标注的语料库，可以为计算机词典提供

有关词的搭配和格框架信息，以及它们在不同类型文本中的分布信息。

3. 语义标注

语义标注的具体内容取决于不同的语言层次。首先可以对文本中的每个词的词义进行标注，其本质是依据上下文甄别句子中每个多义词的确切词义。所以更确切地说，语料库的这种语义标注应该叫做词义标注，或词义排歧(Word Sense Disambiguation，简称 WSD)。其次也可以标注文本中每个句子的句义。比如，用基于格语法的语义网络来表示每个句子的逻辑语义，或只用一对概念结点和它们之间的语义格关系所组成的三元组来分别表示句子中每个意义片断的意思。值得注意的是：美国微软研究院的研究人员正是用语义三元组为细胞，通过对两部英语词典和一部百科全书的句法－语义分析，构造了一个庞大的语义网络 MindNet[67]。MindNet 目前已应用于词义排歧、句法结构排歧和信息检索等一系列自然语言处理领域的研究与开发项目中。

在话题分析中，概念的表达方式可以反映出包含在文本中或谈话人之间的意识形态。在医生和病人的交谈中，一个医生可能使用更多的像“腹部”那样的技术名词，表现出较正式的态度；而另一位医生可能会使用更一般化的像“肚子”这样的词，以便与病人的知识联系起来。另外，在信息检索中，对于一个研究时尚的人来说，如果他想从报纸上了解有关人们在裤子着装上的变化，当他面对一个语料库时，他所要查找

的不仅仅是关键词"裤子"的内容，而且还应该包含"短裤"、"紧身裤"、"牛仔裤"、"马裤"等等。这就是所谓多词一义的问题。也就是说，有多个词同时指向一个概念。此外，在信息检索中，还需要解决一词多义的问题。如果现在想了解在有关物质材料方面的变化，就很可能以"材料"为关键词，但是"材料"还另有一个表示文档的意义。重要的是，人们希望检索到的信息只是对他们有用的信息。

解决上述问题的需求是对文本进行语义标注的动机之一。即给文本中的每个词(次)标注一个词义代码，说明它在当前上下文中所代表的具体意义。根据上述实例，这个代码应该代表由具有同一概念的、意义关联的若干词组成的某个义类。

在进行词义标注时，首先要选择一个语义(或概念)分类系统。在选择分类系统时应该考虑以下的因素：

(1) 最好是在语言学或心理语言学中有共识的分类系统。

(2) 覆盖一种语言的所有实词，而不只是其中的一部分。

(3) 便于灵活修改，以适用于不同用途和不同领域的应用。

(4) 意义的"颗粒度"大小适当。

(5) 分类体系应该具有层次结构。

(6) 分类系统应该遵循一个统一的标准。

4. 话语标注：文本中参指关系的标注

与前面介绍的几种标注类型不同，话语标注是一个比较难以定义的概念。在标注一个文本的话语信息时，可以标注语句这样的单位，它们属于语法的最高单位，在那里句子被标注，并且可以根据句子的话语功能将句子聚类；还可以利用诸如“中心”、“主位”和“述位”等概念来标注这类信息结构；另外，一种话语标注也可以是用结构衔接或参指关系来解释的话语标注。

这里强调的文本结构衔接关系，是迄今为数极少的话语标注实践中，曾经加以实施的一种，也是迄今对一定规模的文本曾经做过的一种话语标注。结构衔接关系同其他如词法、句法和语义的标注不那么紧密相关。词义标注表述了词的意义，而话语标注表述的是语言所具有的极其有效和多方面的能力——把意思从文本的一部分传递到另一部分。如果不能解释这方面的意思，就不能解释人类语言的这部分意思。

会议DAAR'96曾集中讨论了这个主题，代词、参指和参指(歧义)求解构成了不仅是理论语言学，而且是计算语言学的一个主要研究领域。尤其是计算语言学，开始密切关注这一领域中用于训练和软件(或算法)测试的语料库。在过去的几十年间，在机器翻译和信息抽取的领域中，参指求解是最棘手的问题之一。例如，找出第三人称 it, they, he, she, 在一个有待处理的文本中的所指。为解决这一问题，存在两种相反的观点：一种观点认为除非拥有语言的和真实世界的知识，否则机器不可能找出这类参指所指的事物；另一种观点则

认为，利用经验方法可以达到这个目标，这种方法在很大程度上没有真实世界的知识，而仅仅利用句子中参指与先行词之间的距离，以及词或字符等的统计信息。利用统计方法和一个适当标注的语料库，就有可能经验地测试仅仅依靠文本纯形式方面的知识究竟能不能自动辨识先行词。

英国兰开斯特大学受IBM公司的资助编制了一个"参指树库"。这是一个带有句法标注的树库，在此基础上添加了指明结构衔接关系的话语标注。标注实践证明，话语标注可以以一种相对一致和准确的方式来实施。下面给出他们在进行话语标注时的几个实例：

例1：(6 the married couple 6)said that<REF=6 they were happy with<REF=6 *lot*.

例2：(7 this week's winner 7)said <REF=7 he had rung(8 <REF 7 his wife 8)and<REF=7,8 they had spoken to<REF=7,8:2 each other.

其中，先行词用圆括号标注，并且给一个索引号，它在文本中是唯一的；替代词则用"<REF=索引号"来指明其参指关系，即显示该参指词的先行词。

第4章 基于语料库方法的语言学研究

语料库语言学，顾名思义，就是利用语料库方法来研究某些语言学问题。从方法论上讲，这是一种经验的(empirical)方法，以区别于乔姆斯基的唯理的(rational)方法。因而人们自然十分关心在语料库的基础上做过哪些有影响的语言学研究。其实，从实际的语言事实出发是我国语言学家历来遵循的优秀传统。只是由于研究人员在知识结构上的缺陷，使得我国在利用计算机语料库来从事语言研究方面暂时落后于西方。

第一节 语言研究中的语料库方法

本章介绍语料库方法在语言研究中的应用。语料库方法

在语言研究中的一个重要作用就是可以为研究者提供更一般的、经验的语言数据，这些经验数据可以使语言学家作出的结论更客观。本章先介绍语料库方法在语言研究不同领域中的应用，之后给出已经取得某些成果的实例。

一、词汇研究中的语料库

词典编者利用经验数据的历史，远早于语料库语言学的出现。例如，约翰逊（Samuel Johnson）曾利用文学的例句来编辑他的词典。在 19 世纪，《牛津英语词典》（Oxford English Dictionary）利用引用卡片（citation slips）来研究和说明词的用法。收集引文的方法仍在延续，但语料库已改变了词典编者和其他对语料库有兴趣的语言学家考察语言的方式。

语料库意味着，词典编者现在可以坐在一个计算机终端前，仅用几秒钟的时间就可以从数以百万计的文本中调集一个词或短语的全部用法实例。这不仅意味着词典可以比以前更快的生产和修订，从而提供有关语言的最新信息；而且意味着词条的定义能够更完善和准确，因为被观察的对象是一个语料库，是一个比原先规模大得多的语言样本集。

从语料库中抽取出来例子，也可以组成对词汇分析更有意义的组别。比如，按字母顺序对某个词右边的语境进行排序，从而有可能同时观察到这个词的一种特定搭配的全体实例。此外，词典编者拥用的语料库可逐步包括大量的文本信

息，如文本的作者、性别、出版日期和文体，甚至有语言标注的信息；如词性标注、词义标注等。这种能够对信息进行分类的能力，有利于帮助词典编者确定在典型的领域、文体等等具体情况下一个词的各种用法。

通过在同现词语之间建立关系的计算工具（如 3.3 节）来收集词语组合，就有可能比以前更可靠地来考察和处理短语和搭配。熟语（phraseological）单元也许构成一个技术术语或成语，而搭配是认识特殊词义的重要线索[68]。在文本中识别这些搭配，意味着像个别词语一样，它们可以较好地在词典中或机器可读的术语库中得到处理，以提供给专业翻译者使用。

同时基于语料库的词典编纂，可以帮助词典编者从语料库中构造定义。如利用常用的搭配把个别词义捆绑在一起，这使词典编者能将逐词索引数据有效地组织成为词义组，更容易地提供词义频率一类的信息。

二、语料库与语法

和词汇研究一样，语法（或句法）研究也是基于语料库的语言学研究中最常见的实例。语料库对句法研究的重要性表现在：(1)语料库作为整个语言的代表性数据的潜力；(2)作为经验数据，这种语言事实的观察是可以定量统计的。

在 20 世纪 80 年代之前，语法的经验研究不得不主要依靠定性分析。这类研究可以提供语法的详细描写，但很大程

度上难以超越频度和罕见度的主观臆测。随着带句法标记的语料库的出现,以及检索工具的不断开发,使得语法的计量分析(quantitative analysis)比以前更容易进行了。语法的计量分析,至少可以向研究者提供有关什么用法最典型,以及各种变异发生的程度等信息。这不仅对理解一种语言的语法本身是重要的,而且对研究各种不同的语言变异和语言教学也是重要的。

利用语料库的大多数小规模语法研究已经包括了计量数据分析。如施米德(Schmied,1993)[69]对关系子句的研究,给LOB语料库中的关系子句提供了许多计量信息。语料库方法还可以帮助统计各种句型的频率。

50年代以来,语言学家曾分裂成两大阵营:一部分语言学家遵从语言理论的理性主义(rationalist)观点;另一部分则实行描写的经验研究,重视语料库中全部数据的计量。但是,这两派语言学家并不像有些人所说的那样,是相互排斥的。实际上存在一些研究者,他们利用语料库来测试本质上属于理性主义的语法理论,而不是利用语料库做纯语言描写或理论的归纳生成。

在Nijmen大学,理性主义和经验主义两种语法研究被结合起来,以形成一种基本上为理性主义的形式语法。然后以机储语料库所包含的真实语言来测试这些语法。通过参照语言学家的内省,以及对这些语法的解释,首先设计出一部形式语法。之后这部语法被装入一个计算机句法分析器,用它来

分析一个语料库，以便测试它在多大程度上可以解释该语料库中的数据。在这个语料库句法分析的实验基础上，该语法被修正以解释在分析中被遗漏或搞错的语言事实。

另一种利用语料库进行句法研究的思想，就是通过带句法标记的语料库，利用统计数据，归纳出实用语法的类型。有关这种方法的实例见4.2节。

三、语料库与语义学

从前面的章节中，已经知道可以用一个语料库来观察个别词的出现，以便确定它们的词义。这种方法主要用于词典学领域。但更一般地说，语料库在语义学中也有重要意义。它的主要贡献在于为语义学提供了一种客观地解释不确定性和渐变性的方法。

语料库在语义学中的第一个重要作用在于它可以为词项赋义提供客观的准则。语言学家明特(Mindt,1991)[70]曾指出，在语义学中词汇项和语言结构的最高频的意义是通过语言学家自己的直觉来描写的，这是理性主义的方法。但是，语义的辨识事实上必然伴随着文本中可观察的语境—句法的、词法的和韵律的特性，使用语料库可以获得对一种语义差异的客观证据。

语料库在语义学中的第二个重要作用就是更坚定地建立模糊范畴(fuzzy category)和渐变性的思想。在理论语言学

中,范畴通常被认为是一种硬性的分类。也就是说一个词目要么属于某个范畴,要么不属于。但是分类心理学研究认为:认知的范畴通常是非硬性的,而具有模糊边界。因此,问题不在于确定一个给定项是否属于某一特定范畴,而在于它落入该范畴、而非另一范畴的概率有多大。而只有拥有真实的语料库,这种概率的信息才有可能得到。

四、语用学和话语分析中的语料库

迄今,在语用学和话语分析中基于语料库方法的研究还很少。主要原因在于,语用学和话语分析依赖于句间的上下文。语用学常被称为"语境中的意义",而语料库恰恰丢弃了言谈的许多语境。这是由于语料库倾向于采集较小的文本样本,而不是整个文本。而且,在采集到的样本中又进一步删除了它们的社会和文本方面的语境。

国外,语用学及其相关领域的大部分研究都集中于口语。伦敦-隆德(London-Lund)语料库可能是迄今公认的唯一真正的对话语料库。因此,在这方面的大部分研究都是在这个语料库上完成的。这类研究的主要贡献是搞清楚对话是如何进行的,尤其是词、短语和对话之间的关系。如斯特恩斯特朗(Stenstrom,1987)[71]从事的基于语料库的研究,对自发的、自然对话的样本提供了定量的类型学解释。例如,她在观察像带"right"的后续信号时,发现"all right"通常用来在话语中作

为两个阶段之间的一个边界;"that's right"最常用作一个强调信号;而"it's right"和"that's right"用作致歉时的响应。这些计量方法增加了人们对语言行为的理解。因为它们提供了更专门的解释,包括在哪些语境下言谈者可以有哪些选择,在这些选择中哪些是最典型的,或最不寻常的。

五、语言和语言学教学中的语料库

在语言和语言学教学的资源和实践中,通常反映出语言学在经验主义和理性主义方法之间的分野。一方面,许多教科书只采用生造的例句,它们的描写明显地基于直觉或二手材料。另一方面,在柯林斯(Collins) COBUILD 课题组出版的工具书中所展现的例句,则是采用经验主义方法搜集的。他们的例句和描写都依赖于语料库或其他活生生的语言数据资源。

在语言学习中,语料库是例句的重要来源。因为它们在学习的早期阶段就向学生展示了真实的句子和词汇,这是他们在阅读这种语言的真实文本或在真实的交际情景中将会遇到的。这种经验数据,在语言学教学和外语教学中同样重要。但是语料库相对于其他经验数据来源,在语言教育方面有着更重要的作用。它的意义超出了简单地提供真实的语言使用实例的范围。不少学者曾经利用语料库数据来批判现有的语言教学材料。他们采用的方法在很大程度上相似:对教科书

中的相关教学内容或词汇表，利用英语标准语料库，如LOB语料库和伦敦－隆德(London-Lund)语料库进行分析。然后比较他们在这两组数据之间的发现。大多数研究表明，在教科书的教学内容与本族语者如何实际使用这种语言之间，存在很大差别。有些教科书掩饰了语言用法的某些重要方面或用法的多样性。有时甚至牺牲更常用的体裁，而把不常用的体裁选出来并放在突出地位。明特(Mindt)和肯尼迪(Kennedy)等著名学者的一般结论是：没有经验数据为基础的教学材料肯定是误导的。应当用语料库研究指导教材的编写，从而对用法更普通的语言事实给予更多的关注。

外语教学中还有一种特殊的教学类型，叫做“特殊用途的语言”。这是指为一指定应用领域所用的专门语言教学，一般称为专业外语教学。比如，医学专业学生的医学英语教学。建立不同专业的语料库，可以利用它们从事特殊目的的语言教学。香港科技大学曾建设过一个100万词次的英语语料库。样本选自计算机专业学生所用的教科书。这类语料库可以用来提供教学用的许多专门领域的材料。包括词汇表和语言用法的定量数据。这些材料服务于这一特定领域学生的特殊需求，它们比那些从一般化语言语料库中获取的材料更直接。

第二节　现代汉语句型统计与研究

这是北京语言文化大学赵淑华教授主持的一项国家教委博士基金项目，1995 年 4 月 10 日在北京通过了专家鉴定[72]。

1. 汉语句型统计的主要研究目的是：

(1) 通过对中小学教科书和北京语言文化大学对外汉语教材的 400 万字次语料进行句子切分，归纳出现代汉语的句型系统。据此统计出各类句型在语料库中的出现频率，从而为对外汉语教学实践、教材编写、测试标准的制定以及汉外对比等多项研究工作，提供可靠的科学数据。

(2) 为语言学家、语言教学工作者提供一个包含必要数据和句例的熟语料库。

(3) 在中文信息处理方面，将为词组边界自动识别、句子成分自动切分和自动句法分析等研究提供基础性资源。

2. 该课题所取得的具体成果如下：

(1) 对上述 400 万字次语料作了句子切分，除直接输入计算机以外，还制成按上位句型分类并注明出处的句例卡片约 20 万张。

(2) 在上述工作的基础上，对其中的小学语文课本 28 万字次语料作了上位、中位、下位等三个层次的句型分类与统

计。同时对切分出来的每个单句都做了句法分析，并将结果输入计算机，形成了现代汉语句型语料库。总共储存单句16297个。该句型语料库除可向用户提供各类句型的结构特点及相应句例外，还可提供下列数据：

a. 能分别充任主、谓、宾、定、状、补等六大句子成分的各类词语的使用频率，以及定语、状语和补语的语义指向。

b. 出现在谓语部分和宾语部分边界上的词语类型及频率。

c. 多项状语和多项定语的排列顺序。

d. 多项定语中"的"字的分布，多项状语中"地"字的分布。

e. 动态助词"了、着、过、来着"等的分布。

f. 带谓宾的动词列举，等等。

(3) 对北京语言文化大学现代汉语精读课教材34万字次语料的句子作了部分句法分析，摘选了难句，并对部分难句进行了句法、语义、语用三个层面的分析。对部分特殊句式和部分特殊动词作了使用频率的统计。

(4) 对小学语文课本中出现的句型进行了分类，研制了《现代汉语句型使用频率统计表》和《现代汉语常用句型表》。这两个表对研究汉－外句型对比、制定汉语水平测试标准和编写对外汉语教材都有重要的参考价值。例如，过去在对外汉语教材中讲述情态补语(即带"得"字的补语，

又称程度补语)时,一般都要说明这种句子的相应句型:“主+动+宾+重复的动+“得”+补”,如“他写字写得很快”。其实这种句型在现实生活中十分罕见,在小学语文课本的28万字次语料中未见一例。因此,课题组认为在对外汉语教学的基础阶段是否有必要介绍这种句式,是值得重新考虑的。

3. 在这项研究中,课题组对什么是句子作了如下的规定:

(1) 需有完整的句法结构形式;

(2) 能表达相对完整的语义;

(3) 有成句的语调。

4. 在对语料进行句子切分时,他们遇到大量的复句。他们对复句是这样处理的:

(1) 复句中的分句,凡能独立成句的,按单句处理(忽略句中的关联词语);

(2) 如果其中只有一个分句能够独立成句,那么该句作单句处理,其余不能独立成句的按不完全句处理;

(3) 复句中任何一个分句都不能独立的,仍保留为复句。

语料中的复句,不完全句及紧缩句均不进入句型统计。在全部小学语文课本语料中,共切分出单句14387句,紧缩句132句,不完全句578句(如“傍晚回到家”,“张开嘴正要吃”,“他抬头一看”等),复句1200句(如“过了一道湾,又过一道湾”,“他一会儿弯弯腰,一会儿压压腿”,“只要命令一下,他们

就按动板机"等)。

课题组所研究的句型限定为句子的结构类型。区分句型的主要依据是句子构件的性质和它的结构形式,包括句子中的词序、句子成分的多少以及充任某些句子成分的词性等。

5. 随着研究工作的深入,他们对句型的划分又作了如下两点重要的补充:

(1)虽然对句型的划分主要以句子的结构特征为依据,但也不排除在必要时参照句子内部的语义关系。譬如主谓谓语句,虽然同是"大主+小主+小谓"这种格式,但由于大主、小主与小谓的语义关系不同,却可以进一步划分出五种不同的下位句型。如:"大主+小主+小谓(动词/动词词组)"可根据语义关系进一步分成:

SP05 大主语是受事

SP06 小主语是受事

SP07 小主语是大主语的所属部分

SP08 大主语为处所词

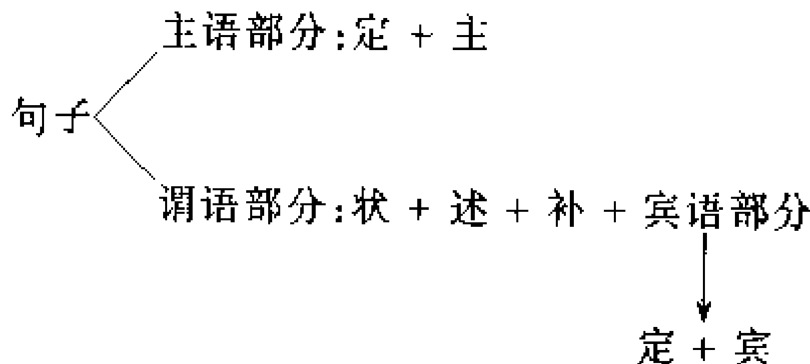
SP09 小主语是对大主语的任指。

(2)按照惯例,定语不是句型成分,因为它的有无不会影响句子的基本结构。但在实际操作过程中,他们发现有的句子缺少了定语便不能独立成句,如:"叶公看见了龙"。但如果加上定语就完全不同:"叶公看见了一条真龙","卖完了好大一叠报","受了一场虚惊"等。因此,定语同状语一样,对于区别句型也是有意义的。因此,对于这种句型的格式就应该确定为:

"主+动+'了'+定+宾"。又如"有"字句,"那姑娘有一双漂亮的大眼睛",其句型格式应该是:"主+'有'+定(数量词及表示描述的词语)+宾"。因为在这种句子中宾语中心词前面的定语是必不可少的。我们不说"那姑娘有眼睛"。其实,这种句子在语义上并不表示领有,而是强调对宾语的描述,全句意思是"那姑娘的眼睛又大又漂亮"。

课题组在析句方法上,采用了句子成分分析法和层次分析法相结合的方法。在划分句子成分时,他们采用了"定+主+状+述+补+定+宾"这种格式。但各成分之间是有层次关系的,并不是同在一个层面上。这种层次关系可以在析句过程中体现。

6. 一个句子首先分为主语和谓语两大部分。主语部分可由"定+主"构成,谓语部分可由"状+述+补+宾语部分"构成,而宾语部分又可由"定+宾"构成,如下图所示。



(1) 词组可以作为一个整体进入句子,充当句子成分。不过这些词组的内部结构只能作句外分析,因为它与全句六大成分(主、谓、宾、定、状、补)之间的结构关系不在一个层次上。

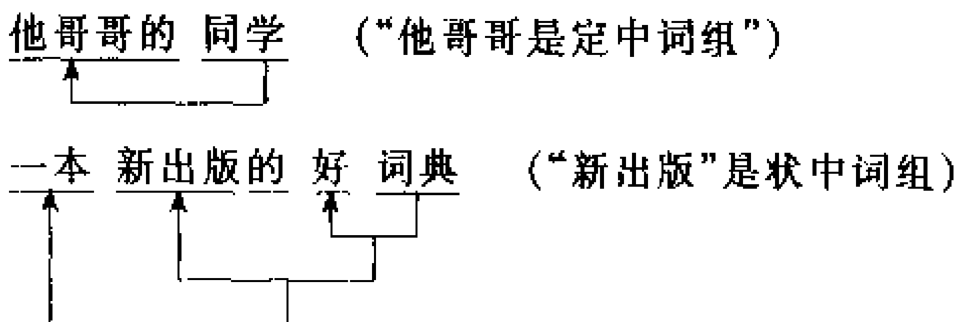
(2) 凡定中词组(在句中是"定+主"或"状+述+补+宾"),其内部层次都是以中心词为起点,自右向左结合,从小到大逐层扩展的。凡述补、述宾、述宾补或述补宾词组,其内部层次都是以中心词为起点,自左向右结合,从小到大逐层扩展的。依循这样一条基本规则,句子的层次关系就可以在析句过程中充分体现出来。

下面以一个动词谓语句为例,说明上述析句过程:

例句:他哥哥的同学昨天在书店买到一本新出版的好词典。

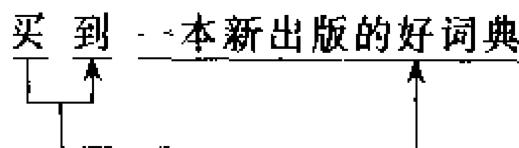
第一步:首先找到谓语部分的中心词——述语"买";

第二步:根据宾语部分和谓语部分的边界,找出句首和句末的两个定中词组,然后以中心词为起点,自右向左结合,逐层扩展。如:

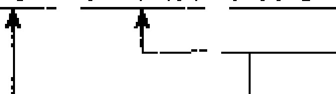


至此主语部分和宾语部分分析完。

第三步:以谓语中心词述语为起点,自左向右结合,逐层扩展。如

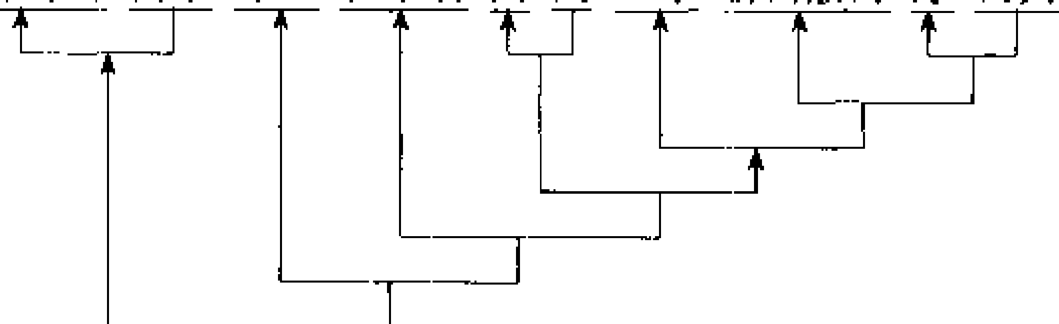


第四步:以述补宾组为起点,自右向左结合,逐层扩展。如:



至此谓语部分分析完。

第五步:将主语部分与谓语部分相结合,即成为句子。如:



第三节 词典学研究

词典学(lexicography)是应用词汇学的一个分支,研究词典编纂的原则与实践,即通过对词项(lexical item)的搜集、比较、注释和分类,进行辞书的编辑。词典学关注的主要问题是词语的意义与用法。不像西方语言学的其他许多领域,在词典编纂领域中历来沿用的都是基于语料库的经验方法。早在1755年约翰森(Johnson)就依据语料库来收集词语的真实用法,把它们作为他主编的英语词典的例句。在18世纪晚叶,《牛津英语词典》(Oxford English Dictionary)也以它所收集的大量文本作为编写的基础,有800多名志愿人员通读这些文本,然后把指定词项的例句写入引文条(citation slips)。然

而这类工作跟今天的语料库方法大不相同。首先,这些早期的文本语料库并不能作为一种语言的代表,而且这些志愿者所关注的往往是词项的特殊用法,而不是它们的通常用法。其次,当今计算机技术的进步,为基于语料库的词典学研究提供了原先意想不到的优势。例如可以设计出在语言整体上更具代表性的语料库,比如包含足够数量的口语材料,具有收集、存储和管理极大规模语料库的能力,以及提供便捷、可靠的语料检索、统计工具等等。和人工相比,计算机可以轻易地在一个几千万词次的语料库中无遗漏地找出任何一个词项在整个词料库中的全体实例,并穷尽地生成这个词项的引文或例句。此外,计算机有能力采用比人工复杂得多的方式来分析任何一个词语的关联模式。例如,当一个词在语料库中有成千上万次的出现时,如果要用人工去计数,并对在该词左右两边 4 个词的范围内出现的其他所有词作出分类和统计,几乎是不可想象的。现在利用一部计算机,这样的任务可以在几分钟内完成。

因此,在更大规模语料库的支持下,利用计算机有可能对词典学研究的一系列问题作出更具代表性、更彻底和更复杂的调查研究,这是过去用手工方式所不能实现的。

早在 60 年代,美国 Heritage 出版社为编纂 Heritage 中学生词典,专门设计了 AHI 语料库[73]。为了使该语料库能收集与统计到真正的学生用词语,即具有代表性,语料库的采样是通过精心设计的。编者首先通过广泛的严密调查,确定了

美国中学生必读或最爱读的1045种正式出版物，然后从中随机抽取1万个样本，每个样本不少于500词次，总库容量达到500万词次。此外，这些语料分属22种题材，从而兼顾了语料的通用性。在此基础上，Heritage出版社不仅出版了Heritage中学生词典，而且在1971年出版了由卡罗尔(Carrol)等人主编的AHI语料库词频统计结果[74]。

被誉为全世界第一部用计算机编纂的词典——《柯林斯COBUILD英语大词典》[75]，也是在规模为2000万词次的COBUILD(Collins Birmingham University International Language Database)语料库的支持下，于1987年完成的。整个计划由柯林斯出版社资助，词典的主编和语料库的设计者是英国伯明翰大学的辛克莱(Sinclair)教授。COBUILD语料库的选材原则也充分考虑了语料的代表性和词典编纂的要求，具体表现为如下六方面：

(1) 语料中书面语占75%，口语占25%；

(2) 语料理应是广为流传的“标准英语”，不收方言。语料中，英国英语占70%，美国英语占25%，其他英语地区的英语占5%；

(3) 反映1960年以来当代英语的用法，收录材料尽可能新；

(4) 不收诗歌、戏剧和科技方面的语料；

(5) 只收作者年龄在16岁以上的成年人作品，并且女性作者的比例不低于25%；

(6) 入选的语料不是样本或片断，而是平均长度为7万词次

的全文或长篇选录,以利于在篇章层次上进行语言分析。

这部词典的出版代表了词典出版业的一个里程碑,其主要特点是在选词、用法和释义等主要环节上都以翔实和定量的语言事实为依据。主编辛克莱声称,这部词典杜绝了编者生造的例句,它选用的全部例句都来自真实语料。仅此一条,就开创了辞书编辑界的一代新风,具有深远的影响。

1997年台湾学者黄居仁、陈克健和赖庆雄主编的《国语日报量词典》[76],是在汉语词典编纂方面首次采用语料库方法的范例。这部词典分两部分:量词部分和名词—量词搭配部分。量词部分说明每个量词的用法,以及可以搭配的名词类型。名词—量词搭配部分是从名词出发,标示能够同指定名词搭配的所有量词。下面是从名词—量词搭配部分随意选出的一个实例:

法

[1] 方法。

方法、办法、作法、手法、用法、写法、疗法、玩法、演算法……。指方法或方式。

[一般] 个、项、套。[种类] 样、式。

看法、说法、想法、讲法……。指意见。[一般] 个、项、点。[种类] 派、样、式。

【辨析】我们可以说"这一点看法、这一点说法、这一点想法",但是不能说"这一点讲法"。

[2] 规律。

宪法、劳动法、刑法、民法、交易法、选举法、国安法、著作权法、保育法、国际法、军法、税法……。指各种法律。通常不配量词。

【辨析】“宪法”还可以说“一部宪法”。法律条文的内容是依“条、项、款”编列，如“民法第一百八十条第一项第二款、公司法第四百一十九条第一项第五款”。

语法、文法、句法……。指语文的规律。

[一般] 套、条、个。

[3] 能力。

枪法、剑法、箭法、刀法、指法……。[一般] 套、条、个。[种类] 派、式。

佛法、魔法……。通常不搭配量词。

书法。[一般] 幅、张、篇。

【辨析】“书法”除了和上述量词搭配之外，还有“他的这一手书法写得真好”这样的说法。

在台湾中研院语料库支持下，这部名量搭配词典的编纂步骤是：

(1) 收集量词、名词词条和直接来自语料库的名词-量词同现数据；

(2) 根据频率对上述数据进行分类；

(3) 由语言学家和词典编者在上述分类数据基础上作出概括。

因此，它的最大特点是：

(1) 不以参考其他词典为主要依据；

(2) 不单凭编辑人员个人的语文素养来编写词条；

(3) 直接从语料库中获取大量的相关例句，再由语言学家加以分析归纳。

这使得这部词典不但更确实地描绘了语言的使用情况，而且更深入地解释了每条词语的用法。比如，每个词条下都罗列了一些最常用、最具代表性的例词和例句。和以往词典不同的地方在于：这些例词和例句不是编辑人员臆造的，而是从大规模语料库中依据使用频率筛选出来、再经编辑润饰而成的。因此，例子生动，变化多，有利于读者举一反三。

具体来说，采用语料库方法可以开展如下几方面的词典学研究：

(1)一个词有几种词义？

这是传统词典学关注的焦点。但语料库语言学使之能通过真实语料的大规模调查来考察个别词项在不同语境中表现出来的相同或不同的词义，而不仅仅依赖于编者的经验或直觉。

(2) 一个词的出现频率。

这类调查就是所谓的词频统计，它可以为我们揭示一个词的通用程度，帮助我们区别常用词和罕用词。这种信息在决定词典的收词、编写语言教科书和开发自然语言处理系统的机器词典等诸多方面都有重要的参考价值。

(3) 一个词通常与哪些词同现？

这就是词语搭配的研究。语言学家弗思(Firth)有一句名言："观其伴，而知其意"[77]。意思是：一个词的词义只能通过与之相伴出现的搭配词才能加以辨识。从这一观点出发，无论是要识别一个词的不同词义(问题(1))，还是学会这个词的不同用法，都必须普遍调查词语的搭配关系和用法模式。统观90年代先后出版的英语词典，几乎毫无例外地把它们的编纂基础建立在大规模语料库的调查之上。不仅如此。朗文(Longman)出版社最新出版的两部词典：《朗文当代英语词典》(第2版)[78]和《朗文英语联想活用词典》(Longman Language Activator)(1993)[79]，都特别关注词语搭配在语言解释和生成两方面所起的显著作用。对那些把英语作为第二语言来学习的读者来说，这一编纂方针尤其重要。因此，编者不仅为一些常用词收入了大量搭配实例，而且干脆打破常规，把大量具有固定用法的短语直接作为词项收列，称之为短语词(phrase word)。在计算语言学和自然语言处理的研究中，词义排歧(Word Sense Disambiguation，简称WSD)是一个公认的难题，而词语搭配的大规模调查，也为这一难题的解决创造了必要的前提。

(4) 语域(register)、历史时期和方言等非语言因素是如何影响一个词的用法模式的？

这类调查有助于搞清楚在不同语域中词语使用的差异，或词语随时间的发展。

下面介绍比伯(D. Biber)对英语词项DEAL(用大写字母

表示它是该词各种曲折变化的总称)所作的几项语料库调查结果[80]。

一、DEAL 的词频调查

从 100 万词次 LOB 语料库中获得的 DEAL 的频率见表 4-1:

表 4-1 DEAL 等词项在 LOB 语料库中的出现频率表

DEAL	290	Deal	182
OF	35749	Dealing	52
THE	2817	Deals	25
HE	9068	Dealt	31
I	7778	总计	290
MAKE	2417		
SIGH	16		
APPROACH	185		
LOOK	500		

DEAL 作为单数或复数名词时在 LOB 语料库不同语域上的频率分布如下表 4-2 所示:

表 4-2 在 LOB 语料库中,DEAL 在不同语域中的频率分布表

语域	分类语料的近似词次数	DEAL 的原始计数	DEAL 的归一化计数(每 10 万词次的出现次数)
新闻报导	88000	14	15.9
述评	34000	4	11.8
社论	54000	4	7.4
宗教	34000	5	14.7

(续表)

学术	160000	16	10.0
民俗	88000	11	12.5
信函	154000	24	15.6
一般小说	58000	5	8.6

由于各种分类语料的容量有很大出入,因此单凭 DEAL 的原始计数还不能直接比较它在不同语域中出现频率高低。所谓归一化,就是把原始计数统一折合成每 10 万词次语料中的出现次数。设原始计数为 m,分类语料的规模为 M,则实行归一化后的计数 n 可用以下公式计算:

例如,LOB 语料库中新闻报导类的 $n=\frac{m}{M}\times 10^5$

语料规模 M=88000(词次),则 DEAL 的归一化计数为:

$$n=\frac{14}{88000}\times 100000=15.9$$

从这项统计中可以看到,在 LOB 语料库的 8 类语料中,DEAL 的出现次数在 5 次以上或 5 次以下的各有 4 类;在信函类中 DEAL 的出现次数最高,也只有 24 次。这说明,100 万词次的 LOB 语料库对于 DEAL 作为名词的用法调查来说仍然显得太小。尽管如此,归一化计数仍然显示出 DEAL 的名词用法在不同语域中确实有不同的分布,比如信函,新闻报导和宗教等三类的归一化计数都大约为社论类的 2 倍。

表 4-3 的统计结果是根据一个更大的语料库——选自 Longman Lancaster 语料库的 400 万词次样本,在小说和学术两类语料上对 DEAL 分别作为名词和动词所获得的频率计

数。

表 4－3 在 Longman Lancaster 语料库中，DEAL 在不同语域中的频率分布表

语域	分类语料的近似词次数	DEAL 的原始计数		每百万词的归一化计数	
		名词	动词	名词	动词
小说	2000000	217	127	107	63
学术	2000000	149	355	74	176
总计	4000000	366	482	90	119

这张频率统计表显示出一个非常重要的事实。尽管对总样本的归一化计数表明 DEAL 作为动词仅比作为名词略为常用一些（每百万词次 119 对 90），但在小说类中名词用法的频率明显高于动词用法（107 比 63），而在学术类中恰好相反，DEAL 作为动词的出现频率达到了名词出现频率的 2 倍以上（176 比 74）。

上述观察结果说明，根据语料总体（总样本）得出的一般化结论，往往不适合某个特定语域的情况。反之，根据某个特定域得到的某种结论，既不代表其他语域的语言事实，更不能推广成为一种语言的普遍规律。换句话讲，词语的出现频率和用法模式在很大程度上是依赖于特定语域的。因此比伯认为，对一种语言（比如英语）所作的某种总体概括往往是误导的，因为它们对不同语域间语言使用方面的重要差异作了平均化处理。一方面，对于一种语言变体（variety）来说，总体概括通常是不准确的。另一方面也可以说，这类总体概括所描写的是一种实际上根本不存在的语言。

二、词义调查

词义调查通常可以通过逐词索引(Key Word in Context,简称KWIC)表入手来做,这种索引表可以穷尽地提供一个被调查的词项在一个语料库中所有的出现及其相伴随的语境。但是对一个多义词来说,要以人工方式去识别被调查词(或称目标词)在每个索引行(或例句)中的词义,也是一件令人生畏的艰难任务。例如,DEAL在一个1000万词次的语料库中大约有2000次出现。对一个更常用的词,检索出来的记录将数以万计。要从中寻找其词义模式,几乎是人工不能胜任的。因此,比伯选择了通过搭配来调查词义的另一途径。所谓搭配词就是经常与目标词在文本中同现的那些词。调查词义分布的这种方式是以下述假设为基础的:一个多义词的每一组搭配词将只与这个词的一个词义发生关联。因此,通过对一个词的最常见的一些搭配词的分析,可以有效地识别该词的一个或多个词义。

表4－4显示了DEAL作为名词的一些最常见的搭配词。所用的语料选自Longman Lancaster语料库的两种语域:学术270万词次,小说300万词次。左搭配词是指那些出现在目标词紧前面的同现词,如“good deal”中的“good”;右搭配是指出现在目标词紧后面的同现词,如“deal of”中的“of”。

表4－4显示,在学术类文本中,名词DEAL的最常见左

搭配词是"great"(每百万词次出现45次),其次是"good"(23次)。事实上,这也是名词DEAL在这个语域的196次出现实例中的185次搭配情况。以下的左搭配词是"package"和"that",分别只出现过2次,或每百万词次平均出现0.7次。

表4-4 DEAL的常见搭配词

语域		搭配词	原始计数	每百万次词次
学术(270万词次)	左搭配	Great	122	45
		Good	63	23
	右搭配	Of	106	39
		More	18	7
		In	8	3
		To	8	3
小说(300万词次)	左搭配	Great	122	40
		Good	84	28
		The	24	8
		Big	10	3
	右搭配	Of	84	28
		To	22	7
		About	15	5
		More	10	3
		With	9	3

这个调查结果还表明,名词DEAL在学术类语料中最常见的搭配是"good/great deal",deal的意思是"数量"或"交易"。如果再考察一下最常见的右搭配词则是"of",每百万词次出现39次,它比下一个最常见的右搭配词"more"(每百万词7次)有明显的优势。于是可推论,作为名词的DEAL来说,最常见的搭配模式是"a good/great deal of",可见最常用

的词义是“数量”。进而把上述观察结果同索引表结合起来，就可以进一步证实对名词DEAL的最常用词义所做的判断。例如索引表显示，“good/great deal”的最常用的用法是“a good/great deal of work”和“a good/great deal attention”。不仅如此，其他几个常见右搭配也显示名词DEAL最常用的词义是“数量”。比如，右搭配词“more”的实例为“a great deal more tolerance”和“a great deal more inhibited”。右搭配的“in”和“to”与“deal”连用时，名词DEAL的词义还是“数量”，如“a great deal in common”，“differ a great deal in their understanding”，“a great deal to be desired”，“a great deal to offer”等。结论是在学术类文本中，名词DEAL的绝大多数出现的词义是表示“数量”。

拿小说类和学术类语料来做一个对比，名词DEAL的搭配情况既有有趣的相似，也有一些明显的差异。一方面我们看到，最常见的左搭配词仍是“great”和“good”。事实上，在这两种语域中搭配模式“good/great deal”的出现频率是一样的，都是每百万词次约68次。另一方面又要注意，在小说类语料中均有96例不能用“good/great + deal”来解释，这些左搭配词分布甚广，如“the”每百万词次出现8次，“big”每百万词次3次，另外还有7种左搭配词只出现过1次或2次。

这说明，名词DEAL表示“数量”的词义在两种语域中都是最常用的词义，但在小说类中出现了许多相对常用的词义（或用法）是学术类中未曾观察到的。例如，左搭配词“the”与

DEAL 联用时，DEAL 有“协议”的词义，如“part of the deal is ...”，“Isn't that the deal ?”左搭配词“big”与 DEAL 联用时，DEAL 的词义是“不重要”，如“no big deal”和“what's the big deal?”

此外，许多未收入上表中的低频搭配词，涉及 DEAL 的另一个重要词义——“交易”，如“property deal”，“record deal”，“cash deal”，“land deal”等。

在小说类文本中，DEAL 的右搭配词同学术类文本的情况十分相似。但“about”和“with”未曾在学术类中出现。

当“about”和“of”与 deal 联用时，DEAL 的词义仍是“数量”，如：

“I also knew a great deal about love.”

“We both laughed a great deal about this.”

而“with”与 DEAL 联用时，DEAL 的词义是“协议”，如：

“I made a deal with the doctors.”

“I'll cut a deal with you.”

在小说类文本中，还发现了 DEAL 在学术类文本中完全没有出现过的一个词义。即在右搭配词词表中原始计数为 4 次的“table”和 1 次的“box”，它们与 DEAL 联用时，DEAL 的词义是“木材”，虽然这两个词都不属于常见的搭配词，但它们却反映出名词 DEAL 在小说类文本中的另一种用法。

比伯把他的调查结果同几本流行的词典作了对比。词项 DEAL 在有些词典中只列出 1 个词条(entry)，而在另一些词

典中最多被列出为4个词条。在这些词条中释义的数目从两三条一直到二三十条不等，读者从中很难推测出DEAL最常见的词义究竟是什么。以下是通过5部词典的调查，从中归纳出作为名词DEAL的重复出现最多的7种词义：

(1) 大量，很多；

(2) 协议；

(3) (纸牌戏中)发牌；

(4) (受到的)待遇；

(5) 分发的行为；

(6) 冷杉木或松树；

(7) 买卖的行为；交易。

绝大多数词典包括了全部7条释义，但5部词典中有2部词典未列出“分发的行为”，一部词典未提及“协议”的意思。释义的排列顺序也存在很大的差异。例如，“大量”的词义出现在韦氏(Webster)词典第一个词条的第二条释义中，而在兰登(Random House)词典中却被排列在第21条释义中。

通过对比，比伯的评论是：

(1) 用DEAL表示“数量”，在被考察的语料库的两种语域中都是最常见的词义，但这个词义在某些出版的词典中却排列得很靠后，分别是第16和第21条释义。

(2) 通过搭配分析找到了名词DEAL的一个较常用的词义，即用“big deal”表示“不重要”，迄今未发现这个词义被上述那些词典收容。

(3) 所有被考察的5部词典都确认"发牌"是名词DEAL的一种词义,它却未曾在比伯采用的语料库中出现。一方面,语料库调查说明DEAL的这一用法是罕见的。另一方面,由于英语的本族语话者(native speaker)都认同这是DEAL的一个独特的词义,把它的释义收入词典是完全正当的。从这个意义上来看,这也可以说是语料库方法的一种局限性。由于语料库采样的偏颇和规模的受限,这类缺陷是很难避免的。因此,在词汇调查中,语言学家的干预也是必要的,他们的学识和语感有时可以弥补语料库方法的不足。

三、搭配研究的语料库方法

搭配在语言教学,尤其是第二语言的教学和机器翻译、语言生成等领域显得特别重要。为什么我们说"穿衣"、"戴帽",而不说"穿帽"、"戴衣"?为什么汉语里可以说"看电影"、"看球赛"、"看小说"、"看朋友",而在英语译文中词"看"必须分别译作see/go to, watch, read和visit?这些都是正确使用一种语言所必须掌握的搭配知识。

《BBI Combinatory Dictionary of English》(John Benjamins Publishing Co.,1986)[81]的作者本森(M.Benson)给搭配所下的定义[82]是:

"搭配是一种强制性的重复出现的词语组合(A colloca-

tion is an arbitrary and recurrent word combination)。"

本森的定义指出了搭配的两条最重要的性质，即强制性和重复出现。

所谓强制性就是必须区分词间的"粘着组合"(bound combination)与"自由组合"(free combination)。粘着组合体现了搭配的强制性。换言之，这种搭配具有一定的特异性，组合中至少有一个词在与其他词组合时受到较大限制，是不自由的。如组合 commit murder(谋杀；犯罪)中，动词 commit 作为"犯(罪)"，"做(错)"解释时只能同屈指可数的几个名词构成述宾关系，这些名词如 crime(罪)、suicide(自杀)、wrong-doing(坏事、恶行)等。因此，commit murder 应看作粘着组合。粘着组合(即搭配)是习惯使然，为什么非得这么讲，不是用句法、语义知识就能解释清楚的，因此是强制性的、不可预期的。反之，自由组合中的每一个词都可以同该组合以外的其他词自由结合，构成相同的句法结构。比如同样是述宾结构，condemn murder(谴责谋杀)属于自由组合，因为 condemn 可以带许多名词(abduction, abortion, abuse of power, acquital 等)作宾语，而名词 murder 也可以出现在数以百计的动词(abhor, accept, acclaim, advocate 等)之后。因此，这种组合不具有特异性，以英语为第二语言的学习者只要掌握了这些词的词义、语法属性及相应的语法规则，便可以在语言交际中根据需要自由生成这样的组合。从这个意义上讲，自由组合不具有强制性，是可预期的。

在汉语的搭配研究方面，国内业已出版了若干部搭配词典。研究汉语搭配，也不可避免地会遇到搭配与非搭配之间界限不清的问题。郭茜认为，《现代汉语实词搭配词典》[83]是现有汉语搭配词典中比较好的一部。但这部词曲仍收录了为数相当可观的自由组合作为搭配实例。该词典的主编在其序言中介绍了他们的编撰思想，即对每个入选的实词都分别进行如下的考查：(1)能否作主语？如果可以，哪些词能作它的谓语？(2)能否作谓语？如果可以，哪些词可分别充当它的主语、宾语和补语？……他们把每个词形象地比喻为一个磁体，被这个磁体吸引的相邻词都被认定为该词的搭配。可以想象，这样收集到的"搭配实例"肯定包括了众多的自由组合(即非搭配)。如主－谓"搭配"："经理能干"、"工人能干"；谓－宾"搭配"："称赞小伙子"、"称赞学生"等等。

因此，在搭配研究中如何引入适当的计量方法来判别搭配与非搭配，就成为语料库语言学的一个重要课题。国外较早利用计算机对搭配进行计量分析的是 Y. Choueka[84]。他们把搭配定义为重复出现的相邻词串，在《纽约时代周刊》(New York Times)约 1100 万词次的文本语料库中，用计算机自动提取出几千条英语常用搭配，如 fried chicken, home run, Magic Johnson 等。这项研究的缺点是没有考虑搭配中两个词可能被其他词隔开的情形(如 make...decision)，以及搭配的强制性要求。丘奇(K. Church)等定义搭配为一组相互关联的词对，他们用信息论的"互信息"(mutual informa-

tion)概念来评价任意两个词的结合能力[85]。他们在一个规模为 4400 万词次的新闻语料库(AP Corpus)上进行实验。互信息这个统计量在相当程度上体现了搭配的上述两个基本性质——强制性与重复出现,并且在计算中对搭配中的两个词是否紧邻没有作出限制。该方法的弱点是未顾及搭配通常具有一定的(句法)结构这一特性,导致他们从语料库中提取到的许多词对,如 doctor-nurse, doctor-bill, doctor-hospital, 虽然在词义上是关联的,但由于不存在句法上的制约关系,因此严格来讲并不是通常意义上的搭配。施马嘉 (F. Smadja) 设计的 Xtract 系统是迄今有关搭配的计量分析中最新和最完整的工作[86]。施马嘉不仅提出了自己对词之间结合强度的计算公式,而且引入了词的位置及其分布离散度的计算公式。Xtract 系统在一个规模为 1000 万词次的股市新闻语料库上提取搭配实例,据称准确率达 80%左右。

清华大学孙茂松等利用 1990 年和 1991 年的新华社新闻语料库 XH-CORPUS,对汉语搭配进行了计量分析的尝试,目的是为语言学家提供搭配的量化依据,以人机交互方式实现搭配的半自动发现[87]。

下面介绍孙茂松实现搭配计量分析的几个量化指标和相关的搭配调查结果。

【搭配强度】

丘奇等曾用互信息 mi 来度量任意两个词 w 和 w, 之间的关联程度:

$$mi(w, w_i) = \log_2 \frac{p(w, w_i)}{p(w)p(w_i)} \qquad (4-1)$$

其中 $p(w, w_i)$是在指定上下文范围内词 w 和 w_i 的同现概率，$p(w)$和 $p(w_i)$分别是 w 和 w_i 在语料库中单独出现的概率。

假设 w 和 w_i 是一个候选搭配对，则式(1)在相当程度上反映了搭配的强制性和重复出现的基本性质。

[1] 在 $p(w, w_i)$ 数值不变的条件下，词 w 和 w_i 之间受约束的程度愈深，它们与其他词同现的机会愈少，因而 $p(w)$或 $p(w_i)$的值减小，从而使 $mi(w, w_i)$的值增大。这表明 w 和 w_i 构成的组合具有较强的强制性。反之亦然。

[2] w 和 w_i 同现的次数愈多，$p(w, w_i)$的值愈大，$mi(w, w_i)$的值也随之增大。这表明 w 和 w_i 重复出现的趋势增强。反之亦然。

在搭配实例的发现中，应把 w 和 w_i 同现的上下文视为同一个句子。在一个句子的范围内，允许 w 和 w_i 被任意多个词隔开。比如在"穿衣服"、"穿新衣服"、"穿了一件红衣服"等短语或句子中，词"穿"和"衣服"都应视作同现。当然，两词相距愈远，一般来说它们发生关联的可能性也愈小。在 Xtract 系统中，把一个英语词的上下文限定存在该词前后出现的各 5 个词。换句话讲，观察窗口的大小是 ± 5 个词。孙茂松在汉语搭配调查中沿用了与此相同的观察窗口长度。

令 $p_j(w,w_i)$ 表示 w_i 在与 w 相距 j 个词的位置上与 w 同现的概率。j = -5，-4，-3，-2，-1，1，2，3，4，5。当 w_i 出现在 w 的左侧，j 取负值；反之，当 w_i 出现在 w 的右侧，j 取正值。

如果搭配强度用 $s(w,w_i)$ 表示，则在式(4-1)的基础上可通过下列公式来计算搭配强度：

$$s(w,w_i)=\log_2\frac{\sum_{j=-5}^{+5}p_j(w,w_i)}{p(w)p(w_i)} \qquad (4-2)$$

假设已知语料库的总词次为 N，若 w_i 在与 w 相距 j 个词的位置上与 w 同现的次数为 $r_j(w,w_i)$，w 和 w_i 在语料库中单独出现的次数分别为 $r(w)$ 和 $r(w_i)$，则用最大似然估计，可分别计算

$p_j(w,w_i)=r_j(w,w_i)/N$

$p(w)=r(w)/N$

$p(w_i)=r(w_i)/N$

把它们代入式(4-2)可得：

$$s(w,w_i)=\log_2\frac{N\sum_{j=-5}^{+5}r_j(w,w_i)}{r(w)r(w_i)} \qquad (4-3)$$

XH-CORPUS 的规模为约 1000 万字次，经自动分词后总词次 N = 7.1 × 10^6，孙茂松在这个语料库上考察了“能力，弱”和“能力，大”两组候选搭配。下面是统计所得的数据。

第一组：“能力，弱”

r_{-3}(能力,弱)=1,r_1(能,弱)=3

r_2(能力,弱)=5,r_j(能力,弱)=0($j=-5,-4,-2,-1,3,4,5$),

r(能力)=2241,r(弱)=177。

根据式(4-3),有:

$$s(\text{能力,弱})=\log_2\frac{7.1\times10^6(1+3+5+7\times0)}{2241\times177}=7.33$$

第二组:“能力,大”

r_{-5}(能力,大)=6,r_{-4}(能力,大)=4,r_{-3}(能力,大)=8,r_{-2}(能力,大)=4,r_{-1}(能力,大)=2,r_1(能,大)=9,r_2(能,大)=6,r_3(能,大)=4,r_4(能,大)=6,r_5(能,大)=5,

r(能力)=2241,r(大)=19913。

根据式(4-3),有:

s(能力,大)=

$$\log_2\frac{7.1\times10^6(6+4+8+4+2+9+6+4+6+5)}{2241\times19913}=3.10$$

从两组候选搭配的搭配强度来看,s(能力,弱)远大于s(能力,大),故“能力,弱”比“能力,大”更像是一个搭配。尽管“能力”和“弱”在语料库中仅同现9次,而“能力”和“大”同现了54次,但由于词“弱”在语料库中单独出现仅117次,“大”则单独出现了19913次,根据式(4-3)的计算,“能力,弱”的搭配强度相比于“能力,大”反而超出。这正与搭配的强制性相吻合。

同理可算出，s(能力，强)＝7.45，s(能力，差)＝6.63，s(能力，小)＝0.74。若按搭配强度的大小排列，有：s(能力，强)＞ s(能力，弱)＞ s(能力，大)＞ s(能力，小)，显示它们成为搭配的可能性依次降低。由于“能力，强”、“能力，弱”和“能力，差”的搭配强度十分接近，数值也比较大，基本上可认定为搭配。s(能力，大)的值陡然下降，它是否应判定为搭配尚有待进一步考察，而 s(能力，小)的值已近乎为零，显然不应视作搭配。

【搭配的离散度】

由于构成搭配的两个词通常还具有一定的结构关系，所以 w_i 在某个或某几个位置上与 w 同现的机会会比其他位置的机会大得多，从而导致 $r_j(w, w_i)$沿位置 j 的分布呈现较大幅度的抖动。而对非搭配来说，$r_j(w, w_i)$的分布则相对平坦。

图4－1显示了两组词“能力，丧失”和“能力，方面”的同现分布：前一组是搭配，其分布变化剧烈，后一组不是搭配，其分布变化和缓。它们的统计数据如下：

第一组：“能力，丧失”

r_{-4}(能力，丧失)＝r_{-3}(能力，丧失)＝1，r_{-2}(能力，丧失)＝8，r_j(能力，丧失)＝0(j＝－5，－1，1，2，3，4，5)

第二组：“能力，方面”

r_{-4}(能力，方面)＝r_1(能力，方面)＝2，r_{-3}(能力，方面)＝3，r_{-2}(能力，方面)＝r_{-1}(能力，方面)＝r_2(能力，方面)＝1，r_j(能力，方面)＝0(j＝－5，3，4，5)

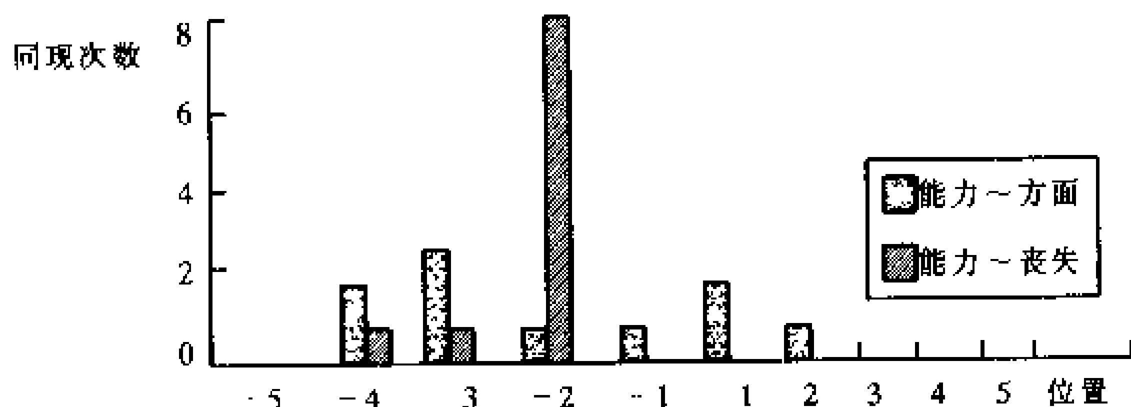


图 4-1 两个词同现的分布

两组词“能力，丧失”和“能力，方面”在语料库中观察到的同现次数相等，都是 10 次。但两组词的同现分布显示出很大的差异。对于 $r_j(w, w_i)$ 分布的离散度 $u(w, w_i)$ 可以用以下公式来计算：

$$u(w, w_i) = \frac{\sum_{j=-5}^{+5}\left[r_j(w, w_i) - \bar{r}(w, w_i)\right]^2}{10} \qquad (4-4)$$

式中，$\bar{r}(w, w_i)$ 是 w_i 在每个位置上与 w 同现的平均次数，显然有

$$\bar{r}(w, w_i) = \frac{\sum_{j=-5}^{+5} r_j(w, w_i)}{10} \qquad (4-5)$$

式(4-4)的分子表示 $r_j(w, w_i)$ 的方差之和。

对上面两组词分别计算其分布的离散度：

$\bar{r}$(能力，丧失) = (1 + 1 + 8)/10 = 1

u(能力，丧失) = ((1 - 1)^2 + (1 - 1)^2 + (8 - 1)^2 + 7 × (0 - 1)^2)/10 = 5.60

$\bar{r}$(能力,方面) = (2 + 3 + 1 + 1 + 2 + 1 + 4 × 0)/10 = 1

u(能力,方面) = ((2 − 1)^2 + (3 − 1)^2 + 3 × (1 − 1)^2 + (2 − 1)^2 + 4 × (0 − 1)^2)/10 = 1.0

当分布的变化幅度很大时,在某些位置上还会出现明显的尖峰,例如图 4 − 1 中"能力,丧失"的分布在位置 $j = -2$ 上,r_{-2}(能力,丧失) = 8,便是明显的尖峰。令 $Z_j(w, w_i)$表示 $r_j(w, w_i)$的 Z − 测试:

$$Z_j(w, w_i) = \frac{r_j(w, w_i) - \bar{r}(w, w_i)}{\sqrt{u(w, w_i)}} \qquad (4-6)$$

则尖峰在位置 j 出现的条件是 $Z_j(w, w_i)$的值足够大。

对图 4 − 1 中"能力,丧失"的分布在位置 $j = -2$ 时,有

$$z_{-2}(\text{能力,丧失}) = \frac{8-1}{\sqrt{5.6}} = 2.96$$

计算结果显示 r_{-2}(能力,丧失)比平均值 $\bar{r}$(能力,丧失)高出 1.96 个标准差,形成一个尖峰。

针对汉语,孙茂松把确定尖峰的算法进一步细化为:

is-peak(w, w_i)

输入:任一词对 w, w_i 在各位置上的同现次数 $r_j(w, w_i)(j = -5, \cdots, 5)$;

输出:是否存在尖峰,以及尖峰的位置;

计算 $\bar{r}(w, w_i)$和各位置的 $z_j(w, w_i)(j = -5, \ldots, 5)$;

对所有的 j,做:

如果 $0.30 \leqslant \bar{r}(w, w_i) < 1.00$，且

$Z_j(w, w_i) \geqslant 2.50$ 或

$1.00 \leqslant \bar{r}(w, w_i) < 5.00$，且

$Z_j(w, w_i) \geqslant 2.00$ 或

$5.00 \leqslant \bar{r}(w, w_i) < 10.00$，且

$Z_j(w, w_i) \geqslant 1.50$ 或

$\bar{r}(w, w_i) \geqslant 10.00$，且 $Z_j(w, w_i) \geqslant 1.00$

则 位置 j 为尖峰位置。

否则 位置 j 不是尖峰位置。

上述算法把 $\bar{r}(w, w_i)$ 划成几个区间，对不同区间上形成尖峰设置不同的阈值 $Z_j(w, w_i)$。这些数据都是通过实验选定的，一般来说 $\bar{r}(w, w_i)$ 愈小，因观察到的同现次数太少，统计数据的可靠性愈差，所以应调高阈值；反之，当统计数据比较充分时，阈值可以适当调低。举例来说，"能力，丧失"在 XH-CORPUS 中仅同现 10 次，$\bar{r}$(能力，丧失)$= 1.00$，数据噪音偏大，故阈值要高些(大于 2.00)，结果只有 Z_{-2}(能力，丧失)$= 2.96$ 可以作为位置 $j = -2$ 处出现尖峰的依据。而词对"能力，大"共同现 54 次，$\bar{r}$(能力，大)$= 5.40$，数据比较可靠，阈值可以低些(大于 1.50)，因此 Z_1(能力，大)$= 1.84$ 也可以作为位置 $j = 1$ 处出现尖峰的根据。说明"能力，大"也可判定为搭配。

离散度与尖峰这两个参数为搭配的计量研究提供了结构方面的判据。孙茂松认为，尽管这两个参数有重要的参考价

值，但并非是判断搭配必不可少的条件。有些搭配，只要搭配强度足够高就行了，离散度不一定要很高，更不一定需要出现尖锋（显然，相对于离散度而言，尖峰是对搭配提出的更“苛刻”的要求）。仅当强度信息不足以作出裁决时，才需求助于离散度和尖峰这两个参数。

以下是孙茂松根据搭配强度（公式 4－3）、离散度（公式 4－4）和尖峰（公式 4－6）等三项指标所设计的搭配判断算法 *is-collocation*(w, w_i)。

输入：任一词对 w, w_i 的强度 $s(w, w_i)$、离散度 $u(w, w_i)$、均值 $\bar{r}(w, w_i)$ 以及各个位置上的 $Z_j(w, w_i)(j=-5, \cdots, 5)$；

输出：w, w_i 是否为搭配。

如果 $\bar{r}(w, w_i)<0.30$，则判定 w, w_i 不是搭配；

如果 $s(w, w_i)\geqslant 4.50$，则判定 w, w_i 是搭配；

否则 如果 $3.50\leqslant s(w, w_i)<4.50$，且

$u(w, w_i)\geqslant 10.00$，

则判定 w，w_i 是搭配；

否则，如果 $2.50\leqslant s(w, w_i)<3.50$，且

$u(w, w_i)\geqslant 20.00$

则判定 w, w_i 是搭配；

否则，如果 $s(w, w_i)\geqslant 2.00$

调用 *is-peak*(w, w_i)；

如果出现尖峰，

则判定 w, w_i 是搭配；

否则，判定 w, w_i 不是搭配；

否则，判定 w, w_i 不是搭配。

在上述搭配判断算法中，搭配的肯定条件有以下三种：

(1) 搭配强度足够大时，无需考虑离散度；

(2) 随着搭配强度的递减，对离散度的要求渐高；

(3) 当搭配强度低到一定程度，必须出现尖峰。

搭配的否定条件也有三类：

(1) 两个词的同现次数太低，数据无统计意义；

(2) 搭配强度较低又未发现尖峰；

(3) 搭配强度太低，即使考虑离散度与尖峰也无济于事。

孙茂松在 710 万词次的 XH-CORPUS 上详尽分析了词"能力"与周围 ±5 个词同现的种种情况。实验结果如下：

在 XH-CORPUS 中，词"能力"共出现 2241 次（即 $r(w)=2241$，w = 能力）。与"能力"在 ±5 个词的上下文范围内同现的不同词共 1932 个，它们均应视作"能力"一词的候选搭配。三条否定条件共排除了 1317 个：其中否定条件(1)排除 962 个，否定条件(2)排除 201 个，否定条件(3)排除 154 个。经三类肯定规则确认的有 615 个：其中肯定条件(1)确认 411 个，肯定条件(2)确认 27 个，肯定条件(3)确认 177 个。这些被确认的词中，包含 88 个根本不可能与"能力"构成搭配的数词（如"一"、"千"）、助词（如"的"、"了"）、连词（如"和"、"无论"）、副词（如"不"、"较"）等，利用一个简单的过滤程序，可将

它们剔除。此外，还有因语料自动分词错误而导致的误判 29 个，如词典中没收“调控”一词，使得自动分词系统把“调控能力”切分为“调/控/能力”，进而导致搭配判断程序把“调 能力”和“控 能力”误码判为搭配。实际上，“调控 能力”才是一个搭配。扣除上述两项因素后，机器最终认定的搭配共 498 个。通过人工审核，其中真正的搭配 169 个。也就是说，对“能力”一词而言，用计算机计量分析方法发现搭配的准确率为 169/498 = 33.94%。表 4－5 给出了“能力”一词的部分实验结果。

表 4－5 部分实验数据(w = 能力 $r(w)$ = 2241)

序号	w_i	$r(w_i)$	$r(w,w_i)$	$s(w,w_i)$	$u(w,w_i)$	尖峰位置	机器判断是否为搭配	机器之判断是否正确	搭配之结构
1	吞吐	78	78	11.30	547.56	-1	是(肯 1)	对	定中
2	分辨	25	3	8.57	0.81	-1	是(肯 1)	对	定中
3	强	1651	91	7.45	138.29	1	是(肯 1)	对	主谓
4	弱	177	9	7.33	2.69		是(肯 1)	对	主谓
5	提高	6058	205	6.75	187.25	-4, 3	是(肯 1)	对	主谓
6	差	638	20	6.63	18.20	1	是(肯 1)	对	主谓 动宾
7	具有	2749	51	5.88	45.49	-3, -2	是(肯 1)	对	动宾
8	和	15377	278	5.84	632.36	-3, -2, 1	是(肯 1)	错	
9	石油	1641	18	5.12	18.20	-2	是(肯 1)	错	
10	有	30354	212	4.47	578.36	-1	是(肯 1)	对	动宾
11	使	8701	62	4.49	57.30	-5, -4	是(肯 2)	错	

(续表)

12	组织	7760	24	3.29	22.64	-1	是(肯2)	对	定中
13	问题	12476	28	2.83	24.16	-2	是(肯2)	错	
14	拥有	1141	5	3.80	0.85	-3	是(肯3)	对	动宾
15	大	19913	54	3.10	3.84	1	是(肯3)	对	主谓
16	不	17409	37	2.75	19.21	1	是(肯3)	错	
17	而	6908	17	2.96	3.01		否(否2)	对	
18	国家	18986	38	2.67	13.16		否(否2)	对	
19	接受	1729	6	3.46	0.44		否(否2)	错	定中
20	要	15404	18	1.89	5.36	-3	否(否3)	对	
21	小	9481	5	0.74	0.85	3	否(否3)	对	
22	活动	7428	8	1.77	2.36	-1	否(否3)	错	定中
23	民族	5473	2	0.21	0.16		否(否1)	对	
24	东	5471	1	-0.78	0.09		否(否1)	对	
25	涵蓄	1	1	11.63	0.09		否(否1)	错	定中

孙茂松把机器发现并经人工核定的搭配同《现代汉语辞海》中词条“能力”下罗列的搭配作了对比，因为这部词典是对现代汉语常用实词的搭配面貌进行尽可能充分描写的巨著。表4-6中，(a)是两者重叠的部分，(b)则是机器通过语料库发现而上述词典没收录的。表4-6是机器发现并且经人核定的搭配的全部清单。

表 4－6　机器发现并经人核定的搭配（共 168 个）

(a)培养 判断 鉴赏 生产 竞争 制造 运输 加工 支付 偿还 平衡 消化 吸收 繁殖 实际 具备 缺乏 提高 强弱 大 差 有限 增强 具有 业务 劳动 适应 领导 组织 分析 保护 发挥 工作 技术 专业 管理 创造 运输 发电 丧失 防御 指挥
(b)反应 扩大 综合 形成 达到 设计 抗灾 开采 影响 排水 客运 保障 承受 一定 执政 反应 安置 配套 不足 超过 出口 自立 创汇 动手 吞吐 增加 运行 足够 防务 操作 处理 作战 通信 同等 自给 自理 防守 减弱 现有 约束 作业 防卫 鉴别 通航 负重 不够 生存 隐蔽 科研 失去 抗病 炼油 腐蚀 后续 识别 抗旱 削弱 限制 识字 存储 自主 对抗 核算 机动 消费 分流 超出 防洪 自卫 干扰 免疫 再生 信任 过剩 供给 应急 饲养 运算 扑救 防疫 驾驭 筛选 参政 相应 采油 整体 通行 核定 载荷 维修 运载 接待 保存 分辨 保鲜 装备 耐寒 通车 转换 防范 自救 联运 决策 独到 起重 输送 新 有 开发 服务 群众 发展 测量 显示 突破 依靠 强化 控制 经营 供应 下降 监督 低 核 拥有

孙茂松对搭配的语料库调查说明：

(1) 强度 $s(w, w_i)$，离散度 $u(w, w_i)$ 和尖峰是搭配计量分析中三个比较合适的统计量，但它们只是一种相对的标准，表 4－5 中在肯定条件下均列有误判的例子；

(2) 统计数据的分布特性相当程度上反映了搭配的结构性质。

图 4－2 显示搭配“能力～具有”呈动宾结构（尖峰位置 －2，－3）；图 4－3 显示搭配“能力～差”呈主谓结构（尖峰位

置+1)；图4-4搭配"能力～提高"，既有动宾结构(尖峰位置-3，-4)，又有主谓结构；图4-5搭配"能力～吞吐"-峰突兀(尖峰位置-1)，反映了汉语中一种很典型的搭配格式——偏正式名词短语。

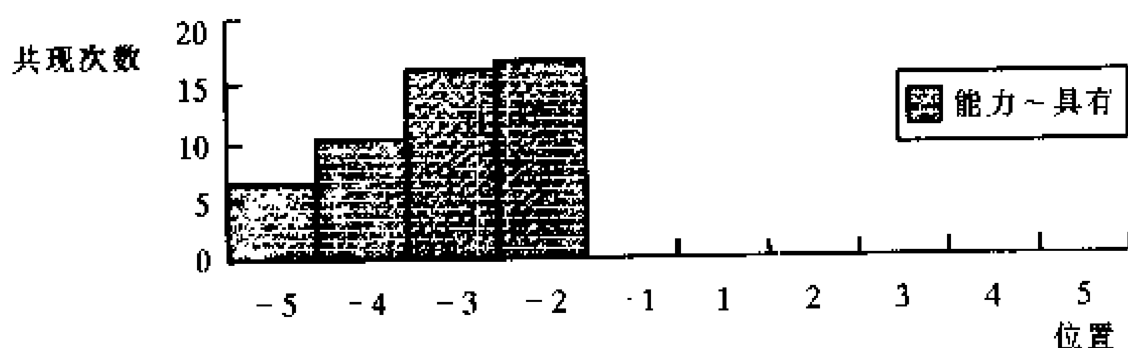


图4-2"能力—具有"的分布

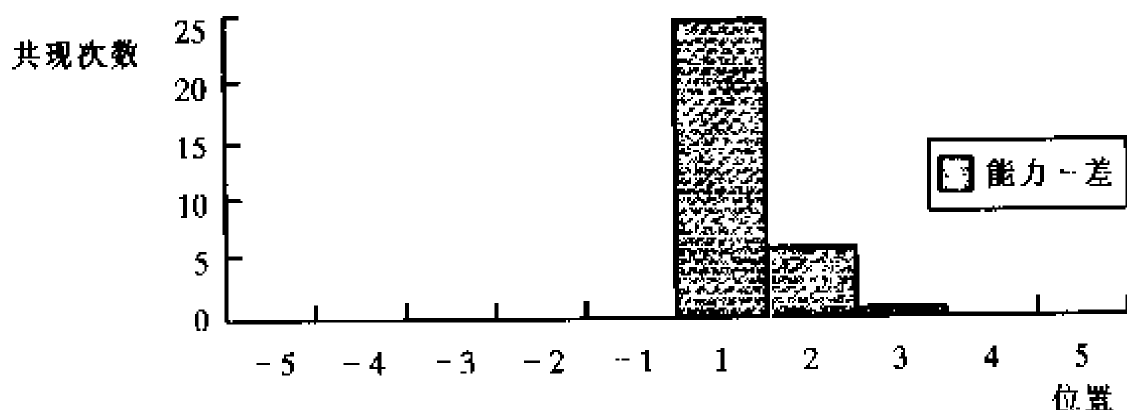


图4-3"能力—差"的分布

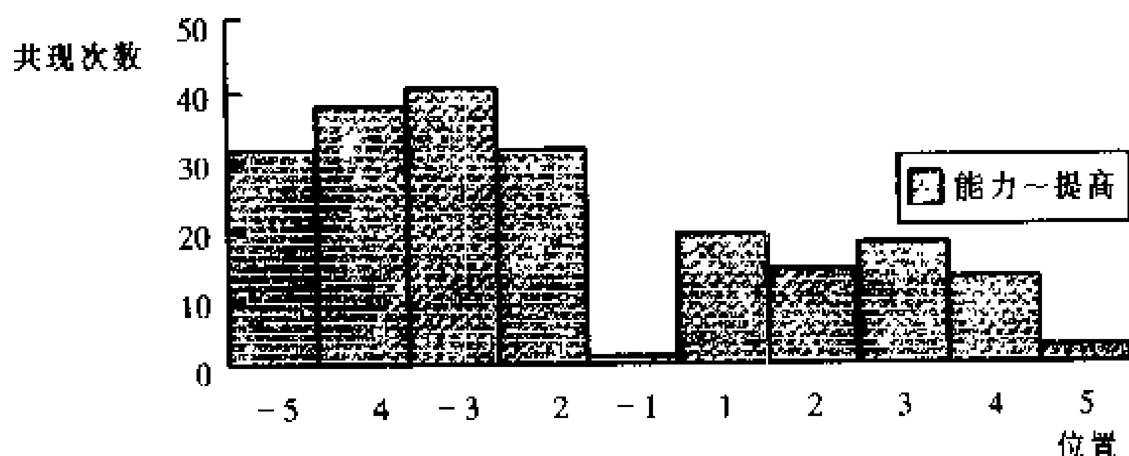


图4-4"能力—提高"的分布

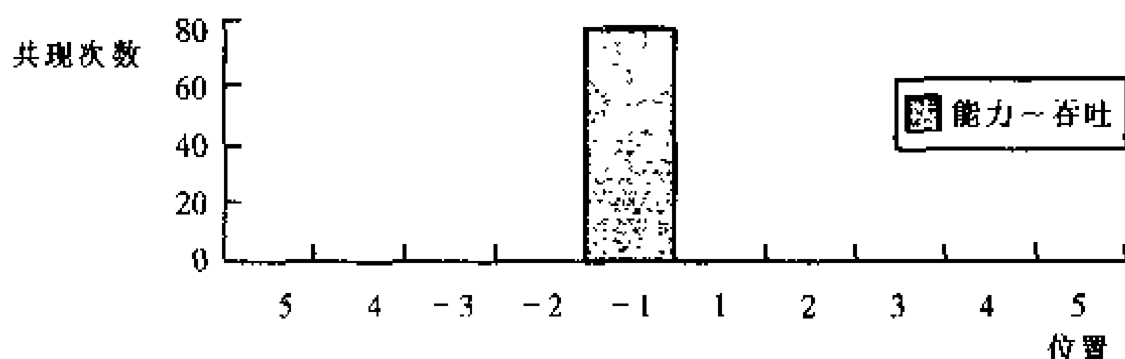


图4-5 “能力—吞吐”的分布

(3) 搭配是受领域制约的。人们大体上认同的搭配如“阅读能力”、“写作能力”,由于领域不同,在 XH-CORPUS 中未曾出现。在给定领域中,语料库的规模和采样情况也会对搭配的计量分析带来重要影响。如表4-5中的“接受能力”(强度3.46)、“活动能力”(尖峰位置-1),“涵蓄能力”(强度11.63,但仅同现1次),人们一般会认可它们是搭配,从统计数据来看,它们也确实“差不多”是够格的,可惜在 XH-CORPUS 中同现次数太低,只能割爱;

(4) 对阈值的设置,应在搭配发现的准确率(指在机器发现的搭配中被人工核定为搭配者所占的比例)与召回率(指语料库包含的搭配总数中被机器判定为搭配者所占的比例)之间寻求某种折衷。一般来说,阈值越高,准确率越高,而召回率越低;反之,阈值越低,召回率越高,而准确率越低。孙茂松的语料库调查期望发现尽可能多的搭配,“宁滥勿缺”,所以把阈值定得比较保守。

针对“能力”一词30%左右的准确率是否偏低了呢?施马嘉曾指出,采用传统办法手工编辑 Oxford English Diction-

ary (OED)的准确率只有大约4%。相比之下，利用计算机来发现搭配比之人工，效率有了大幅度的提高。此外，传统方法容易受到人为因素的干扰：研究者的语感常常是因人而异的，由于不同人的语言、知识背景不尽相同，他们对搭配的判断是主观的、定性的，相互协调也非易事。利用计算机对大规模语料库实行搭配的计量分析，有望减轻语言学家的工作强度，提高搭配的质量与覆盖面。

第四节　汉语名词的语义分类研究

众所周知，名词—量词搭配现象是汉语不同于印欧语言的一个重要特点。多数学者还认为，在现代汉语中量词的主要功能是对名词的语义进行分类。因此通过大规模语料库的调查，掌握名词—量词搭配(同现)的计量数据，有可能揭示汉语名词的语义分类机制[88]。这正是台湾学者黄居仁、陈克健、高照明主持的一项研究工作的宗旨。这项研究直接利用了作者根据中研院语料库调查获取的名词—量词搭配数据。在这些定量数据的基础上，他们用信息熵的公式来计算名—量搭配的信息内容，用向量方式来计算任意两个名词群之间的(语义)距离。论文表明用以上方法可以推导出一个大致可行的汉语名词分类体系。本书选取这个实例来介绍基于语料库的语言学研究，有以下两个原因：

(1) 该研究充分利用了从中研院语料库中直接抽取的名—量词搭配实例及其统计数据，其结果体现于国语日报社 1997 年出版的《国语日报量词典》一书中。

(2) 为了根据量词的搭配情况对汉语的名词进行聚类，就必须分别对量词和名词作出形式化表示，并给出其相似程度的计算公式。在这项研究中，一个量词所携带的信息量(即熵)用它可以修饰的名词数的多少来计算；而每个名词的语义则用它所能接受的那些具体量词(所组成的向量)来表示。因此通过对名词向量之间的距离计算，就可以对全体名词进行聚类，得到名词的语义分类机制。这种对处理对象给出形式化表示，以及给处理方法设计出合理的计算公式，就是语言建模(modeling)的典型方法，具有普遍意义。

为了理解黄居仁等人的研究工作，这里有必要重温一下申农(Shannon)信息论中有关一个随机事件的信息量的概念[89]。

假设 S 代表一组随机事件 $E_1, E_2, \cdots, E_n$，其中任一事件 E_i 出现的概率为 p_i。根据概率的基本知识有：$0 \leqslant p_i \leqslant 1$，$p_1 + p_2 + \cdots + p_n = 1$。

定义事件 E_i 的信息量为

$$I(E_i) = -\log_2 p_i \quad (\text{比特}) \qquad (4-7)$$

根据概率的特性可知，随机事件的信息量 I 恒大于零；

事件出现的概率愈小，它所携带的信息量就愈大。确定性事件($p_i=1$)的信息量为零，也就是说它的出现是意料中的事，没有给我们带来任何信息。

根据系统 S 中每个事件的信息量 $I(E_i)$，可以用系统的熵 $H(S)$来表示 S 中各事件信息量的统计平均值，即

$$H(S)=-\sum_{i=1}^{n}p_i\log_2 p_i \quad (比特) \qquad (4-8)$$

由于一个随机事件的信息量随其不确定程度的增长而加大，所以熵是一个系统不确定程度的度量。熵不可能是负值，熵愈大，系统的不确定程度愈高。反过来讲，如果 S 是一个确定性系统，则它的熵为零。

假设系统中共有 N 个名词，其中 n 个名词可以与指定量词 X 搭配同现，那么这个量词 X 所携带的信息量是以下两个熵值之差：

$$I(X)=H(N)-H(n) \qquad (4-9)$$

如果名词和量词搭配的事件具有相等概率，则具有 N 个名词的系统中每个名词与某一量词搭配的出现概率是 $1/N$，它的熵

$$H(N)=-\sum_{i=1}^{n}\frac{1}{N}\log_2(\frac{1}{N})=\log_2 N$$

同理，对于系统中能够同某一量词 X 搭配的 n 个名词来说，其熵

$$H(N)=\sum\frac{1}{n}\log_2(\frac{1}{n})=\log_2 n$$

所以,量词 X 所携带的信息量

$$I(x)=\log_2 N-\log_2 n \qquad (4-10)$$

由于 N 为常量,因此上式告诉我们:能与之搭配的名词越少,这个量词所携带的信息量就越大,它对名词语义分类的贡献就越大。这似乎同我们的直觉是一致的。

系统中的每个名词(或名词群)用一个多维向量表示,向量的每一维代表系统中的一个具体的量词,向量在某一维上的值等于该量词的信息量(公式4-10)。如果该名词(或名词群)不能同某个量词同现,则其向量的这个分量为零。在向量表示的基础上,两个名词(或名词群)之间的语义亲近程度(affinity)可以通过这两个向量之间的距离来量度。

在上述假设下,这个研究小组通过重复地把两个最相似的名词类结合成一个新的名词类的办法,最终形成一棵名词类语义树。对名词类进行聚类的算法如下:

(1) 利用公式(4-10)计算182个量词的信息量;

(2) 每个量词对应于182维向量空间中的一维,每个量词的信息量对应于一个向量分量;

(3) 每一类名词都对应一个向量,它是用与该类名词搭配的全部量词所对应的向量之和来定义的;

(4) 重复地对具有最小距离的任意两个名词类进行聚类,以形成一个新的名词类,并用一个相应的向量来表示,这个新向量到原先的两个向量的距离相等(即$(v_1+v_2)/2$)。不断地重复这一步骤,直至只剩下单独一个名词类。

在上述第一步中，计算了每个量词的信息量。结果显示，信息量最小的量词是“个”，其信息量为1.269。这是意料之中的结果，因为“个”是公认的最一般化和中性的量词，可以与它同现的名词很多，因此在名词分类中它的贡献最小。接下去，信息量为3.363的是量词“名”，它是表示人类的个体名词的一般量词。信息量与“名”很接近的量词是“位”、“群”和“只”，它们都是可以和大量名词同现的一般量词。在名词－量词系统中信息量最大的量词是“阙”、“题”、“桌”和“班”，它们都只能唯一地同一个名词群搭配，它们的信息量是11.52。

请注意在第三步中，代表每个名词群的向量是用能与该名词群搭配的全体量词来定义的。在语言学上，这种做法可以解释为：用全体可搭配的量词来描写这群名词共享的语义属性。因此，具有相同的可搭配量词的那些名词群之间是不可区分的，它们理应合并为一个名词类，并用一个唯一的向量来表示。所以，尽管在量词词典中有1910个名词结尾和超过2000个词条，在名词—量词搭配系统中只能分出502个不同的名词群和它们相应的向量。

上述聚类过程的结果是一棵只具有单独一个母亲结点的二叉树。树的终结结点是用可搭配量词定义的名词类。任何一个名词类同一个与它距离最近的另一个名词类合并，形成一个新的名词类。该过程重复地执行，直至全体名词类都被聚合到一个母亲结点之下。作者提出的假设和算法是否合理、可信，就取决于我们是否能对这棵树作出合理的语义解

释。

黄居仁等的实验结果表明,在深度小于 4 的子树中可以得到可靠的语义分类结果。他们通过这种办法获得了 50－75 种有意义的名词类,其中两个名词类如图 4－6 所示。

a.房子,屋子［个,栋,间,幢］

宿舍,校舍,房舍,精舍,官舍,公寓［栋,间,幢］

楼房,洋房［个,座,栋,间,幢］

官邸,宅邸,大厦、广厦、华厦、别墅,古厝,大厝,庙宇,寺宇,屋宇,楼宇,宅院［座,栋,间,幢］

镖,飞镖［支,枝,枚］

冲天炮,烟斗,箭,弓箭,利箭,弩箭［支,枝］

b.标杆,竹竿,撑竿,钓竿,鱼竿,矢,箭矢［支,枝,个,根］

扫把,火把,矛,铁矛,长矛,竹蜻蜓,木棍,铁棍,警棍,烟卷［支,枝,根］

长鞭,竹鞭,教鞭,马鞭,烟,香烟,大麻烟,洋烟,长寿烟［支,枝,根,条］

栏杆,电线杆［支,枝,个,根,排］

图 4－6:部分名词聚类结果

实验结果还表明,当树的深度大于 5 时,聚类结果与直觉大相径庭。对此,作者作了如下解释:

(1)用等距离法来表示合并后的新向量,可能是不正确的。这使得某些因素消失得过快。这是因为在计算量词的信息量时,没有考虑非搭配量词其实也会对名词的分类作出贡

献。所以这种方法不能区分语义冲突和语义无关这两种不同情况。从理论上讲,具有冲突的语义属性的两个子类应当互相抵消。这意味着对于描写合并得到的新类来说,这种特定的属性是无关的。然而,如果非同现仅仅是由于跟这个子类无关,那么标注在另一子类上的一个语义属性仍会被合并所形成的类所继承,显然其描写能力略弱。为解决这个问题,需要设计更精细的模型。然而,这类模型将要求在大量词典数据上标明每个名词群不能与某些量词搭配的原因。从方法论上来讲,这将导致不能用经验数据来证实的那种有争议的判断。为此,他们无意在这个方向上进一步深入。

(2)注意到量词本身通常是歧义的。例如,量词"条"可指明下列七种语义属性:(i)细长的物品,(ii)体形呈长条状的动物,(iii)条状地上物(隧道,水道等),(iv)"线",包括抽象意义的线,(v)法律、规则、讯息,(vi)人命,(vii)歌曲。黄居仁等的研究把每个量词当作一个唯一的符号,而不去区别其不同的语义属性。这样做的优点是处理方便,但会把不同的语义属性错误地聚合到同一类,仅仅因为它们共同拥有一个形式相同的符号。黄居仁小组目前正利用来自量词词典和名词-量词搭配词典的信息,以便获取颗粒度更细的搭配关系来解释每个量词的所有词义。这样一来,名词群的初始数从原先的502增加了一倍,达到1061个。原先由于颗粒度粗而被错误地聚合在一起的语义上不同的名词群,现在得以正确地分离。这应导致名词聚类的更好结果。

第五节 词汇—语法问题调查

通过考察与一组同义词关联的不同语法结构,我们可区分极其接近的同义词;而通过考察与同义词的语法结构关联的不同词类,我们可以区分十分接近的语法结构。这类研究被称为词法—语法关联(lexico-grammatical associations)研究。

比伯(D. Biber)曾对两个意义十分接近的同义形容词"little"和"small"进行了语法关联的考察,试图通过不同的用法模式对这两个同义形容词作出区别[90]。调查表明,尽管这两个词的意思都是"小",而且都可以出现在句子的定语或谓语位置上,但是它们在这两个句法位置上表现出明显的偏向性,而且这种偏向性与不同语域紧密关联。

在英语中,定语形容词(attributive acjectives)位于名词前面,用来提供有关那个名词的信息。例如:

"The little girl next door pulled him through the fence."

"But I'm not a small person."

而处在谓语位置的形容词则出现在一个系动词(copula)之后,其功能是向句子的主语提供信息。例如:

"When she was little, she couldn't say Jessica."

"Did you think it would be too small ?"

被调查的语料库包括两个部分:来自英国国家语料库(BNC)的500万词次对话语料,以及来自Longman-Lancaster语料库的500万词次学术语料。整个语料库都经过词性自动标注,其中每个词都注有词性信息,包括是定语形容词还是谓语形容词的标记。

下面是"little"和"small"用作谓语形容词的百分比:

	对话类语料	学术类语料
little	2%	<1%
small	23%	13%

统计结果显示,在两类语料中这两个形容词在绝大多数情况下都出现在定语位置,而不是谓语位置。同时显示,"small"比之"little"有多得多的机会出现在谓语位置上,在对话类语料中更是如此(23%,而学术类语料是13%)。不论在哪一类语料中"little"都极少以谓语形容词出现(对话类中为2%,学术类中不足1%)。

比伯还对两个差不多同义的动词"begin"和"start"进行了语法关联的考察。在大多数情况下,这两个动词可以互相替代。如

"After the race started..."

"After the race began..."

事实上,这两个动词在句法组合功能上具有完全相同的语法潜能,即相同的"配价"(valency)能力。比如,它们都可以

同时充当及物动词和不及物动词，如：

（1）及物模式——取一名词短语作直接宾语：

"Then they started/began [the quota system]".

（2）不及物模式——没有直接宾语：

"I had better issue a survival kit before we start/begin."

在及物模式中，直接宾语既可以是一个名词短语，也可以是一个"to-子句"或"ing-子句"，"begin"和"start"都可以采用后两种变体。如：

（1）及物模式——以"ing-子句"作为直接宾语

"They had started/begun [leaving] before I arrived."

（2）及物模式——以"to-子句"作为直接宾语

下面是比伯在小说（200万词次）和学术（200万词次）两种语域上对这两个动词的语法关联情况所作的调查。全部语料均来自Longman-Lancaster语料库。

表4-7 对"begin"和"start"的语法关联情况调查

		不及物模式	及物模式			总计
			+NP	+to-子句	+ing-子句	
Begin	小说类	54(22%)	8(3%)	178(72%)	10(4%)	250(100%)
	学术类	82(43%)	22(12%)	66(34%)	22(12%)	192(100%)
Start	小说类	101(40%)	55(22%)	50(20%)	44(18%)	250(100%)
	学术类	91(64%)	22(16%)	21(15%)	8(6%)	142(100%)

语料库调查表明，两个动词在两种语域中在所有可能的配价模式中都出现了。但统计显示出两个重要的用法模式：

（1）"start"与"begin"相比，更常用作不及物模式；

(2)“begin”与“start”相比，更常用作带“to-子句”作为宾语的及物动词。

统计结果表明，“start”在小说类语料中的40%出现，在学术类语料中的64%出现是不及物的。相反，“begin”在小说类中只有22%，在学术类中只有43%的出现是不及物的。

通常，在学术类语料中“start”作为不及物动词出现时，主要描写某个过程的开始，如：

“... the soil formation process may start again in the fresh material”.

“Blood loss started about the eighth day of infection ...”

“Tillering starts about a week or earlier after broadling.”

以上例句的主语通常是一个名物化的过程，而且在动词后面跟着出现一个状语。在小说类语料中，句子的主语更多是具体的人和物，然而在动词后面仍经常出现一个状语。如：

“As he started down the hill, he could see it ...”

“... the train had started again ...”

因此，与这些动词的不及物用法一起联用的状语是一个值得进一步研究的课题。

统计还显示，在两种语域上“begin”比起“start”都更常用作及物动词。在小说类文本中，“begin”的72%出现是用“to-子句”充当宾语的；在学术类文本中这种情况也达到34%。

相比之下,"start "用"to-子句"充当宾语的比例仅达 20%(小说类)和 15%(学术类)。

由于对"to-子句"的自动识别具有很高的精度,比伯把上述使用模式在 Longman-Lancaster 语料库的 1000 万词次语料上作了进一步的验证。表 4 -8 就是在更大规模的语料库上得到的调查结果。

结果验证了先前的结论,即动词"begin"的及物性用法与"to-子句"形式的宾语发生强烈关联,尤其是在小说类语域上 60%;"start"则与其他及物模式发生更强的关联。

表 4-8 对"begin"和"start"的语法关联情况在更大语料库上的调查

		+ to-子句	所有其他及物模式	总计
Begin	小说类	1 871(60%)	1 267(40%)	3 183 (100%)
	学术类	221(36%)	387(64%)	608 (100%)
Start	小说类	294 (17%)	1 462 (83%)	1 756 (100%)
	学术类	90 (14%)	541 (86%)	631 (100%)

比伯通过语料库调查说明,尽管"begin"和"start"这两个动词具有几乎相同的词义和完全一样的配价模式,它们在不同语域中的实际使用模式却迥然不同。这说明人们凭直觉(或语感)对于这些使用模式所作出的判断(或指导)是不可靠的。包括本族语的话者(native speaker)在内,人们往往不能正确预见这类系统地关联的用法模式的存在,最多只能预见哪些动词与哪种配价模式关联。与此相反,基于语料库的调查表明,对表面上同义的那些词,如果从它们的用法关联模式

来考察，它们很少是完全等价的。

第六节 语域变体 (register variation)研究

语域是根据它们的使用情景来定义的，如它们的目的、话题、场所、交互性和方式等。对任何一个话者来说，掌握(或控制使用)不同语域是至关重要的。可以毫不夸张地说，没有一个人只会使用单独一种语域。即使在一天之内，人们也会采用不同语域的方式来讲或写同一种语言。因此，一个人需要有能力在不同语域之间作出正确的选择(或转换)。在一个人语言习得的所有阶段上，语域特性的习得都具有重要意义。

不论是为了理解语域习得的过程，还是为了使语言教师们掌握语域教学的有效方法，都需要首先解决不同语域的语言特性的描写问题，以便能正确区分不同的语域。尽管研究人员早已认识到这类描写的重要性，但事实证明：除了语料库方法以外，这一目标是很难达到的。这是因为深入的语域研究，有三个基本要求：

(1)大量文本作为研究素材；

(2)考虑大量的语言特性；

(3)在不同的语域间进行定量的对比。

而这些要求都离不开大规模语料库，以及相应的语料库分析

技术。

首先，占有大量文本是进行这项研究的基础，因为在文本数量不足的条件下得到的语域研究结果不可能是准确的。

其次，以少量语言特性为对比基础的语域研究，不可能提供深入的语域描写。在此基础上作出的概括也不可能准确。事实上不可能只凭单独一个突显的语言特性就能识别一个语域，除非它只出现在这个特定的语域中。实际情况是：许多语域往往共享一批语言特性，如名词、代词、动词和形容词等的出现频率。只有通过这些特性的相对使用频度，才能对这些语域加以辨识。换句话讲，核心语言特性在使用中呈现的系统性差异提供了区别不同语域的可靠依据。

最后，语域分析要求一种对比方法。即需要提出一条比较的基线(baseline)，以便判断在某一种语域中一种(或一组)语言特性的出现究竟有多少。例如，比伯的调查显示：在英语文本中，关系从句在每1000词次中出现25次，就算是极频繁的了，因为它在平均出现频率通常是每1000词次1～10次，具体情况取决于不同语域。相反，对名词来说，大约是在每1000词次中平均出现200次。因此，如果在每1000词次中名词只出现25次，就被认为是罕见的情况了。

下面介绍比伯(Biber)怎样进行口语和书面语两种不同语域变体的研究[91]。这不仅是因为口语和书面语之间的差异一直是各国语言学界的一个热门话题，而且通过这项研究可以看到比伯在区分这两种语域时采用了哪些语言特性，以

及他首创的多维分析(multi-dimensional analysis)方法。

在进行口语和书面语的广泛对比时,要知道哪些语言特性可以用来作为对比的基线是十分困难的。比如通过调查可以看到,在英语中关系从句的出现,在学术文体和对话文体中表现迥异。相反,在这两种文体中都很少使用过去时动词:学术体在每1000词次中平均出现22次,在对话体中每1000词次平均出现35次。但在小说体中,过去时动词在每1000词次中平均出现80次。

鉴于这种原因,不可能只凭个别语言特性的相对分布来识别不同语域。要考虑的语言特性实在太多了,而又不确切知道哪一个特性在所考察的语域中是重要的。其实,研究表明,在对话体中呈现的语言特性包括:块状的(fragmented)句子结构、紧缩形式(contraction)、第二人称代词(you),插入成分(you know),半情态动词(have to, need to, be able to)以及wh-补足语子句等。相反,在学术体中呈现的语言特性有:频繁出现的名词、定语形容词、名物化和其他特殊的词项,被动结构和外置构造(如 it is possible that)等。

尽管一般来说,语言学家已经认识到在特定语域中某些重要的同现模式,但要对这些同现模式进行计量统计是十分困难的。事实上,要准确识别在不同语域上同现的语言特性群,只有采用语料库方法才能真正实现。比伯在80年代提出的多维分析法已证明语料库是解决这一问题的基础。

作为语域变体的多维分析的基础是一个包括广泛样本的

口语和书面语语料库，它必须是全面代表一种语言（如英语）的主要同现模式。在80年代，对英语所作的多维分析，利用了一个综合语料库，它由481个文本组成，约960000词次。其中，340个文本选自LOB语料库，基本上覆盖了主要的书面语语域，如学术、社论、小说等；其他141个文本选自伦敦一隆德（London-Lund）口语语料库，如面对面的对话、公开演说和有准备的讲演稿等。

多维分析的第一步是确定要考查的语言特性集。这一步骤的目的是收集广泛的语言特性，它们应能对文本的功能关联作出相应的解释。比伯列出了在英语多维分析中采用的全部67个语言特性，它们可以大体归入以下16个主要的语法范畴：

（1）时、体标志

（2）处所和时间状语

（3）代名词和代动词（pron-verb）

（4）疑问

（5）名词形式

（6）被动

（7）状态形式

（8）主从关系特征

（9）介词短语、形容词和副词

（10）词汇特殊性

（11）词类

(12) 情态动词

(13) 特殊化动词类

(14) 简约形和不爱用的结构

(15) 并列

(16) 否定

下一步是开发计算机程序,以便对文本中每个语言特性的每次出现进行识别和计数。有些复杂的特性需要采取人—机交互的识别方式。全部标注结果需经人工检查,以保证统计的正确性。

显而易见,对英语语料库所作的上述调查来说,分析员所面对数据量是令人生畏的。语料库包含 481 篇文本,每篇文本有 67 个语言特征的频率统计结果。为了找出文本中那些同现的特征,比伯采用了一种名为"因素分析"(factor analysis)的统计方法。这是一种统计相关(correlational)技术,目的是识别分布方式类似的变量组。即通过因素分析来找出哪些语言特征倾向于在文本中同现。

比伯把每个同现特征群叫做语域变异的一"维"(坐标)。比如,一个同现特征群可以由第一人称代词、第二人称代词和 wh-疑问组成;另一个同现特征群由名词、介词短语和定语形容词组成,等等。通过因素分析获取的这些语言特征群反映了它们各自的元素在不同文本中以相似方式分布的语言事实。例如,当某一篇文本出现大量名词时,一般来说,它也应该大量出现介词短语和形容词;反之,如果在某一篇文本中,

名词的数量很少,它似乎也只有很少的介词短语和形容词出现。

由于语言特征的同现反映了一类文本的共享功能,所以在找到了用来定义"维"的那一组语言特征之后,就可通过情景、社会功能和认知功能等因素来对这一"维"作出功能上的解释。例如,在对话语域中第一人称代词、第二人称代词,直接疑问句和祈使句的频繁出现,可以用对话的交互性来解释。直接疑问句和祈使句的使用,要求有一个听者来对言谈作出响应;而第一人称和第二人称则分别直接指称话者和听者。类似地,紧缩形式、错误话头(false starts)和一般化实词(如"thing")都与实时的言谈约束有关。

比伯在英语口语和书面语的多维分析中识别了变异的5维。下面列出了和第一维相关联的同现特征。每一维由互补出现的两个特征群组成。也即当其中一个特征群在某一文本中频繁出现时,与其互补的另一特征群则出现得特别少,反之亦然。这两个互补特征群分别标以"正"和"负"。在第一维中。正特征群包括:私动词(如"think","feel"),(在补足语子句中)that省略,紧缩,现在时动词,和第二人称代词等。负特征群则包括:名词、词长、介词短语、型/标比和定语形容词等。

第一维:

私动词	0.96	名词	-0.80
that-省略	0.91	词长	-0.58
紧缩	0.90	介词短语	-0.54

现在时动词	0.86	型/标比	-0.54
第二人称代词	0.86	定语形容词	-0.47
……			
可能性情态动词	0.50		
非短语并列	0.48		
wh-子句	0.47		
句尾介词	0.43		

在每个特征右边都有一个数字,表示该特征在第一维中的"负荷"(loading),是该语言特征与这一维之间的相关强度(strength of the relationship)的度量。可以看作是这个语言特征在这一维的各个成员中的代表性(representative)。负荷的取值范围从 +1 到 -1。负荷的绝对值越大,其代表性就越高;当负荷值等于1时,叫做完全相关。可见,在第一维的特征中,私动词和名词(它们的负荷分别为 0.96 和 -0.80)比之可能性情态动词(0.50)和定语形容词(-0.47),是这一维的更强代表。

每个语言特征在任意一维中都有某个负荷值。由于具有较大负荷的那些特征是这一维的更佳代表,所以在关于这一维的功能解释中也更有用。通常仅当一个特征的绝对值大于 0.30 时,才在这一维的功能解释中予以考虑。

在负荷值的基础上,通过对绝对值大于 0.30 的那些特征在一篇文本中出现频率的统计,可以计算出这篇文本在某一维上的得分,简称维分(dimension score)。进而对语料库中

每种语域全体文本进行统计，就可计算出每种语域的平均维分。在这个基础上就可以描写任何指定语域的语言特性，进行任何两个语域之间的对比，进而对每一维作出完整的功能分析。

第5章 语料库方法在计算语言学中的应用

20世纪90年代以来在自然语言处理(NLP)和计算语言学的研究中,语料库方法和统计语言模型迅速崛起,成为主流技术,决不是偶然的。通过以下一些实例介绍,相信读者会从中领会到这种新方法的魅力。

第一节　汉语文本中交集型切分歧义的研究

在自动分词系统中,歧义切分字段和未登录词是影响分词精度的两个最主要的因素。一般来说,歧义切分字段又可进一步划分为交集型及包孕型(或多义型)两类。在真实文本

中绝大多数歧义切分字段属交集型。据国家863智能计算机主题组1995年对国内部分自动分词系统的评测报告[92]，交集型歧义切分的正确率最高为78%，包孕型歧义切分的正确率最高仅达59%。这些数字说明，对歧义切分问题的研究仍将是中文信息界关注的焦点之一。

以下集中介绍山西大学和清华大学分别在大规模语料库上对交集型歧义切分字段所做的普查。为了解释他们的研究成果，有必要先统一说明一下下文中用到的几个基本术语，有关这些术语的形式化定义请参见文献[93]。

【交集型歧义切分字段(简称交集字段)】假设A，B和C分别表示由一字或多字组成的字串。如果在ABC字串中，A，AB，BC和C都是词表中的词，则称ABC为交集型歧义切分字段。当然，交集字段也有比此更复杂的情况，详见下文。如果分词过程仅依据词表来进行，没有其他句法、语义知识，那么AB/C和A/BC都是合法的切分结果。例如，“应用于”可切分为“应用/于”和“应/用于”；“可以为”可切分为“可以/为”和“可/以为”。因此，“应用于”和“可以为”都是交集字段。

【交集因子】交集字段中相互交叉的词叫做交集因子。如交集字段“应用于”中，两个交集因子分别是“应用”和“用于”。

【交集字段的链长】一个交集字段中交集因子的个数称为该交集字段的链长。因此，交集字段“应用于”和“可以为”的链长都是2。显然，一个交集字段至少包含两个交集因子。换句话说，任何交集字段的链长至少为2。同理，每个交集因子的

长度不应小于 2,而任何交集字段的长度不应小于 3。

【两个交集因子的耦合段】两个相邻交集因子的重叠部分称为它们的耦合段。耦合段的字数称为耦合长度。交集字段"应用于"的耦合段是"用",所以耦合长度为 1。

【最大交集型歧义的分字段(简称最大交集字段)】假设 S 为任一字串,且 S 的一个子串 S_1 是交集字段,如果 S 中不存在包含 S_1 的更大交集字段,则称 S_1 是 S 的最大交集字段。下面的例子是一个更复杂的交集字段:



该字段包含 3 个交集因子:"大部分"、"分手"和"手工业品"。前两个交集因子的耦合段为"分",后两个交集因子的耦合段为"手",耦合长度的均为 1。整个交集字段的链长为 3。值得注意的是,该交集字段中"手工业"和"工业品"这两个词也相互交叉。但由于它们被一个更长的词"手工业品"所包含,所以"手工业"和"工业品"不是交集因子。

区分最大交集字段的目的在于:最大交集字段不再跟它周围的任何字串形成新的交集因子,是一个相对独立的字段。这使得我们能撇开该字段的上下文环境,把它分离出来加以单独处理。例如,在以下例句中:

"经 济 法 有 普 遍 的 强 大 约 束 力"。

"强大约"和"强大约束力"都是交集字段,但前者被后者所包

含。所以,在本例中"强大约"不是最大交集字段。只有"强大约束力"在句中不为其他交集字段所包含,成为该句的最大交集字段。

在交集字段或最大交集字段的语料库普查中,分别用到段型比(静态频率)和段次比(动态频率)等两个统计值。它们的定义如下。

【段型比】段型比指某种类型的(最大)交集字段个例数在(最大)交集字段个例总数中所占的百分率。计算公式是:

$$\text{段型比}(\%)=\frac{\text{某种类型(最大)交集字段的个例数}}{\text{(最大)交集字段个例的总数}}\times 100\% \tag{5-1}$$

【段次比】段次比指在一个语料库中某种类型(最大)交集字段的出现次数在(最大)交集字段出现总次数中所占的百分率。计算公式是:

$$\text{段次比}(\%)=\frac{\text{某种类型(最大)交集字段的出现次数}}{\text{(最大)交集字段出现的总次数}}\times 100\% \tag{5-2}$$

一、山西大学的普查

山西大学郑家恒、刘开瑛曾在一个180万字次的新闻语料库上对交集字段进行过一次普查,他们未对交集字段和最大交集字段加以区分[94]。对交集字段链长的定义也同上文略有区别,为了阐述的方便,本文统一用以上定义的术语来介

绍她们的调查结果。

通过语料库调查，郑、刘共取得9500条交集字段个例，并据此建立了一个交集字段库。考虑到不论在词典还是在语料库的统计中，二字词均占总词条数的70%左右。因此，他们首先把重点放在以二字为一切分单位的交集字段上。根据他们的调查报告，在9500条交集字段中，这种以二字为一切分单位的字段就达8378条(88.2%)。表5-1是这8378条交集字段的链长分布统计结果：

表5-1　交集字段的链长统计结果表

链长	交集字段数	段型比	出现次数	段次比
2	4646	55.4%	14581	64.1%
3	3409	47.7%	7632	33.6%
4	231	2.8%	381	1.7%
5	92	1.1%	149	0.6%
总计	8378	100%	22743	100%

从统计数据可以看出，链长为2和3的交集字段数占总数的96.1%，它们的出现次数则占总次数的97.7%。因此，如果能解决好链长为2和链长为3的交集字段切分问题，就能大大提高整个歧义字段切分的正确率。

郑、刘的论文还对不同链长的交集字段分别做了切分结果的调查，并给出了自动切分的相应对策。他们的报告仍限定在以二字为一切分单位的交集字段上。

(1)对链长为2的交集字段ABC，切分结果有如下4种情况：

a.切分结果为A/BC，如“出自己”，切分结果为“出/自

已”；

b.切分结果为 AB/C。如“出现在”，切分结果为“出现/在”；

c.切分结果为 ABC，如“传染病”；

d.切分结果不定。

对 180 万字次新闻语料所作的统计结果见表 5－2。

表 5－2 交集字段 ABC 的切分调查表

切分结果	交集字段数	段型比	出现次数	段次比
A/BC	2330	50.1%	6258	42.9%
AB/C	1827	39.3%	5505	37.7%
ABC	374	8.0%	2604	17.9%
不定	123	2.6%	217	1.5%
总计	4654	100%	14584	100%

统计结果显示，切分结果为 A/BC 和 AB/C 的交集字段数占总数的 89.4%，出现次数达到总次数的 80.6%。其中 AB/C 型在正向最大匹配的切分过程中可以得到正确切分。因此，对链长为 2 的交集字段，切分难点集中在 A/BC 型交集字段的辨识上。

(2)对链长为 3 的交集字段 ABCD 中，切分结果有如下几种情况，详见表 5－3。

表 5－3 链长为 3 的交集字段的切分结果调查

切分结果	交集字段数	段型比	出现次数	段次比
A/BC/D A/B/CD	55	1.6%	103	1.4%

（续表）

AB/CD	3331	97.7%	7489	98.3%
ABC/D AB/C/D	7	0.2%	8	0.1%
ABCD	5	0.2%	5	0.06%
不定	11	0.3%	11	0.14%
总计	3409	100%	7616	100%

统计表明，AB/CD型无论在交集字段个例数还是在出现次数上都占总数的约98%，如"已经过去"切分结果为"已经/过去"。因此对链长为3的交集字段，一般来说切分结果应为AB/CD。

(3)链长为4的交集字段ABCDE中，从切分结果来看，主要难点在前3个字的切分歧义上。如"为人民工作"，只要把前3个字"为人民"正确的分为"为/人民"，就能得到整个字段的正确切分结果："为/人民/工作"。

(4)链长为5的交集字段ABCDEF中，正确切分结果均是AB/CD/EF。如"中国产品质量"，正确的切分结果是："中国/产品/产量"。

根据语料库的统计，郑、刘发现：在4646条链长为2的交集字段中，只有8条在不同上下文中会有不同切分结果，如：

"从小学"：姐妹/二/人/从/小学/上/到/中学。

她/从小/学/戏剧/表演。

"以北约"：确立/以/北约/为/核心/的/军事/力量。

兴平/市/以北/约/十五/公里。

而在链长为3的3409条交集字段中,没有观察到同一字段在不同上下文中出现不同的切分结果。因此,他们认为在研究交集字段的切分策略时,应把重点放在不同的交集字段上,而把极个别的、在不同上下文中存在不同切分结果的交集字段作为个例处理。

在语料库调查的基础上,山西大学的研究小组归纳出了针对链长为2的交集字段切分规则,把这些规则运用到4646条链长为2的交集字段个例上进行封闭测试,切分正确率达到87%。然后,他们又对从2万字次新闻语料中抽取的链长为2的交集字段进行开放测试,切分正确率达到81%。

应当指出,山西大学的上述研究结果是针对交集字段中以二字为切分单位所作的统计,而且采用的语料库规模只有180万字次。交集字段的实际情况会比他们的观察更复杂。

二、清华大学的普查

孙茂松和他的学生利用清华大学的词表TH-WL,内含112967个词条,在一个规模为101506152字次的新闻语料库Rcorpus上,无遗漏地抽取出其中所有的最大交集字段共233888条[95]。这些字段累计在Rcorpus中出现1793317次,字次总数为6566244,覆盖了整个Rcorpus的6.47%。

图5-1给出了前 n 个高频最大交集字段对Rcorpus全部最大交集字段的覆盖率 r 的曲线 $r(n)$。统计显示,前

2500个高频最大交集字段的覆盖率已超过50%，前4619个高频最大交集字段的覆盖率达到59.2%。

为验证上述统计结果，他们又在另一个领域包括新闻、技术、科普、军事的60万字次语料库Acorpus上来考察上述4619个高频最大交集字段的覆盖率。调查结果如图5－2所示。图5－2说明，Rcorpus的前4619个高频最大交集字段对Acorpus中全部最大交集字段的覆盖率仍然达到50.9%。这个数字虽比对Rcorpus的覆盖率59.2%略低，但下降幅度有限。说明在Rcorpus上获取的高频最大交集字段是较稳定的，受不同领域语料的影响不大，具有相当程度的通用性。

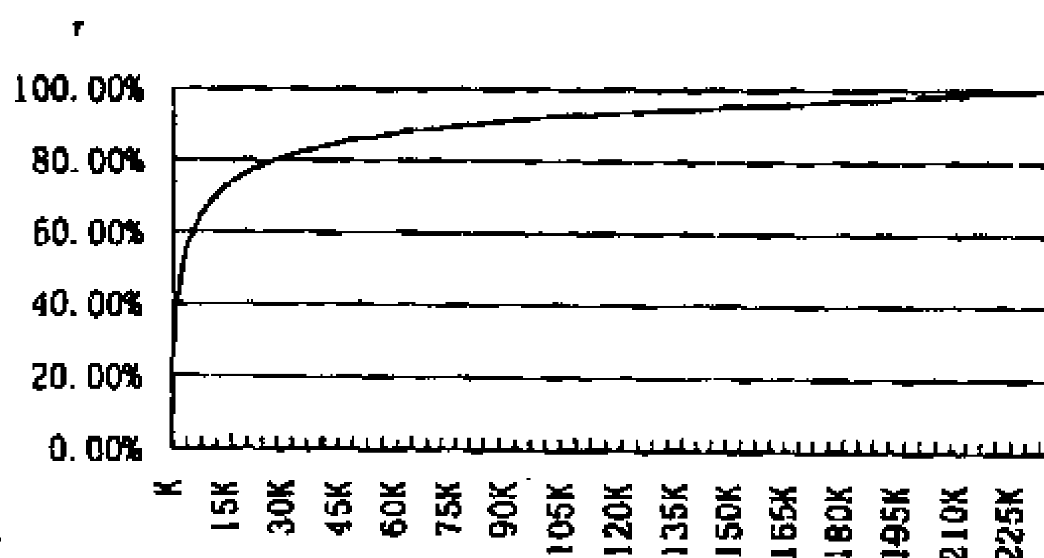


图5－1　在Rcorpus中，

前n个高频最大交集字段对全体最大交集字段的覆盖率r

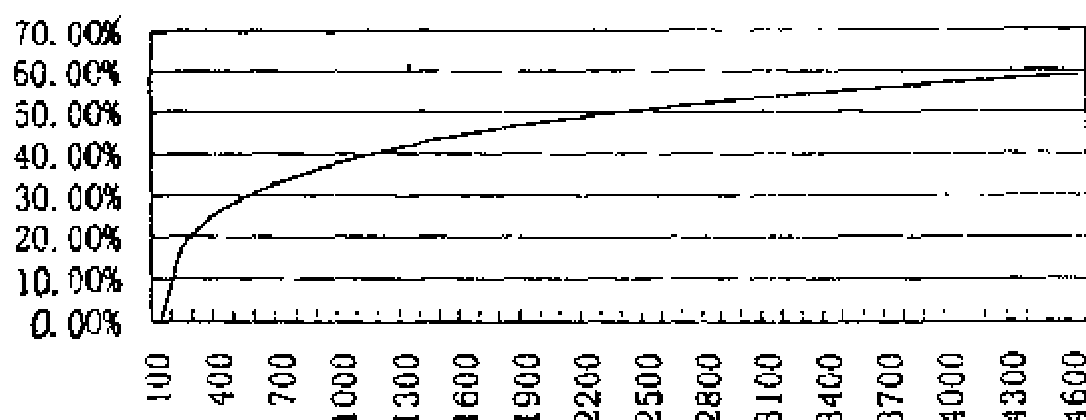


图 5-2　Rcorpus 前 n 个高频最大交集字段关于 Acorpus 的覆盖率 r

孙茂松把 Rcorpus 中前 4619 个高频最大交集字段归纳成以下 3 种类型：

(1)伪歧义：只有一种切分结果；.

(2)真歧义 1：经常存在两种以上切分结果；

(3)真歧义 2：本质上属真歧义，但通常只有一种切分结果。也就是说，其他形式的切分结果出现机会很小，基本上可当成伪歧义处理。

它们的统计结果如表 5-4 所示。

表 5-4　关于歧义类型的统计

歧义类型	最大交集字段数	段型比
伪歧义	4279	92.6%
真歧义 1	85	1.9%
真歧义 2	255	5.5%
总计	4619	100%

伪歧义的突出性质是其切分结果是唯一的，与上下文无关。因此，对伪歧义型高频最大交集字段来说，可以把它们唯一正确的切分结果预先记录在一张表中，对它们的切分只需

通过简单的查表操作即可完成。由于伪歧义加上真歧义2的最大交集字段数已覆盖 Rcorpus 前4619个高频最大交集字段98.1%,和 Rcorpus 上全体最大交集字段的58.6%,所以上述这个极其简单的分词策略,对解决交集字段的切分问题不失为一种十分有效的对策。

考虑到最大交集字段的外部特性,如字段长度、耦合长度和链长等因素,将对交集字段的切分策略直接发生影响。孙茂松在文献[96]中分别对这三个因素在语料库中的分布情况作了统计。统计结果如下:

(1)Rcorpus 中,最大交集字段的长度分布

表5-5中"具体例子"一栏内的数字分别用来表示一个交集因子的起始位置和长度。例如表中第一行的第一个例子:"项目的(0,2)(1,2)",(0,2)表示这个交集字段的第一个交集因子"项目"始于该字段的首字(位置为0),长度为2;(1,2)表示第二个交集因子"目的"始于该字段的第2个字(位置为1),长度也是2。

表5-5 最大交集字段的长度分布表

最大交集字段长度	最大交集字段段型数目	段型比	段次比	具体例子
3	77861	33.29%	49.76%	项目的(0,2)(1,2),上海市(0,2)(1,2),为人民(0,2)(1,2),和服务(0,2)(1,2)

（续表）

4	106256	45.43%	39.89%	在意大利(0,2)(1,3)，离退休金(3,3)(1,3)，行政区域(0,3)(2,2)
5	29481	12.60%	5.98%	进一步到位(0,3)(1,4)，自来水龙头(0,3)(2,3)剩余劳动力(0,4)(2,3)
6	15583	6.66%	3.52%	申请人民法院(0,3)(2,4)，自由市场经济(0,4)(2,4)
7	3055	1.31%	0.61%	少数民族自治区(0,4)(2,4)(4,3)，与此同时差不多(0,4)(3,2)(4,3)
8	1190	0.51%	0.17%	主持人情不自禁地(0,3)(2,2)(3,4)(6,2)，扎扎实实地下功夫(0,4)(3,2)(4,2)(5,3)
9	280	0.12%	0.04%	领导人民建立新中国(0,3)(2,2)(3,2)(4,2)(5,2)(6,3)
10	108	0.05%	0.02%	青年突击队长生龙活虎(0,5)(4,2)(5,2)(6,4)
11	55	0.02%	0.01%	全民所有制表演艺术团体(0,5)(4,2)(5,2)(6,2)(7,3)(9,2)
12	14	0.01%	0.00%	合法政党参与国家政治生活(0,2)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(7,2)(8,2)(9,2)(10,2)
13	3	0.00%	0.00%	乌兹别克斯坦共和国外交部长(0,9)(8,2)(9,2)(11,2)

（续表）

14	2	0.00%	0.00%	提高人民生活水平息息相关(0,2)(1,2)(2,4)(5,2)(6,2)(7,2)(8,2)(9,2)(10,4)
总计	233888	100%	100.00%	

统计结果表明，字长为3和4的最大交集字段的段型比和段次比分别高达78.82%和89.65%，无疑是Rcorpus中最大交集字段的主体。如果把字长为3,4,5,6的最大交集字段都累加在一起，它们的段型比和段次比更高达97.98%和99.15%。显然，这四种字长的最大交集字段应成为研究人员注意的重点。

(2)Rcorpus中耦合段的耦合长度分布

统计结果表明，最大交集字段中相邻交集因子形成的耦合段，其耦合长度绝大多数为1(达到段型比的99.57%和段次比的99.04%)，仅极少量耦合长度为2和3。尚未发现耦合长度在4以上的耦合段。

(3)最大交集字段的链长分布

表5-6 耦合段的耦合长度分布

耦合长度	耦合段的段型数	耦合段的段型比	耦合段的段次数	耦合段的段次比	具体例子
1	404161	99.57%	2607582	99.04%	比例如何(0,2)(1,2)(2,2)， 弄虚作假发(0,4)(3,2)

（续表）

2	1732	0.43%	25212	0.96%	民族资本家(0,4)(2,3)，留洋博士生(0,2)(1,3)(2,3)
3	10	0.00%	42	0.00%	犹如箭在弦上(0,2)(1,4)(2,4)，现行反革命分了(0,5)(2,5)
总计	405903	100.00%	2632836	100.00%	

表 5－7　最大交集字段的链长分布

链长	最大\交集字段段型数目	段型比	段次比	具体例子
2	95148	40.68%	57.18%	表现在(0,2)(1,2)，留学生会(0,3)(1,3)，国民经济基础(0,4)(2,4)
3	115467	49.37%	39.99%	任何时候(0,2)(1,2)(2,2)，革命根据地(0,2)(1,2)(1,3)
4	14610	6.25%	1.81%	中国营养协会(0,2)(1,2)(2,3)(4,2)，我国民族资本主义(0,2)(1,2)(2,4)(4,4)
5	7702	3.29%	0.92%	野外科学工作(0,2)(1,2)(2,2)(3,2)(4,3)，逐出世界杯赛(0,2)(1,2)(2,3)(4,2)(5,2)
6	645	0.28%	0.06%	在野生动植物资源(0,2)(1,2)(2,2)(3,3)(5,2)(6,2)
7	275	0.12%	0.04%	进行经常性爱国主义教育(0,2)(1,2)(2,3)(4,2)(5,4)(8,2)(9,2)

（续表）

8	26	0.01%	0.00%	按时运抵交割地上海(0,2)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(7,2)
9	11	0.00%	0.00%	城乡居民生活水平稳固(0,2)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(7,2)(8,2)
10	2	0.00%	0.00%	各国人民生活水平和美化(0,2)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(7,2)(8,2)(9,2)
11	2	0.00%	0.00%	合法政党参与国家政治生活(0,2)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(7,2)(8,2)(9,2)(10,2)
总计	233888	100.00%	100.00%	

统计结果显示，链长为2和3的最大交集字段占绝大多数，它们累计的段型比高达90.05%，累计的段次比更高达97.17%。统计中观察到的最大链长为11，在1亿字次的新闻语料库中仅见2例。

进一步观察还表明，相同类型的最大交集字段会出现很不相同的内部结构。对于内部结构不同的最大交集字段，显然也就有不同的分词策略。例如，“重工业区”和“棉花生产”都是长度等于4的最大交集字段，但前者的链长为2，两个耦合因子分别是“重工业”和“工业区”，耦合长度为2；后者的链长为3，三个耦合因子分别是“棉花”、“花生”和“生产”，耦合长度都是1。



有些结构的最大交集字段，也许用词类信息就可得到满意的切分结果，但用同样办法来处理其他结构则可能差强人意。孙茂松把最大交集字段的内部结构描写分为宏结构和微结构两种情况，并且分别给出了它们相应的描写方法和统计结果。

(1)宏结构类型的描写方法和分类统计结果

由于交集因子是形成交集字段的基本结构因素，因此可以根据交集因子对最大交集字段 S 进行分类。具体来讲，就是用一对括号中的两个整数分别表示一个交集因子在 S 中的起始位置和长度。例如："重工业区"的宏结构类型记作(0,3)(1,3)；"棉花生产"的宏结构类型记作(0,2)(1,2)(2,2)。

对 Rcorpus 中全部 233888 条最大交集字段按宏结构分类，共得到 312 类，表 5-8 列出了其中 12 种主要的宏结构类型。统计数字表明，最大交集字段的宏结构分布是比较集中的。平均每一种宏结构类型含近 750 条最大交集字段个例，最高频的两种宏结构是(0,2)(1,2)和(0,2)(1,2)(2,2)，它们累计的段型比达 73.63%，段次比达 84.53%。

表 5－8 最大交集字段宏结构的分布

最大交集字段的宏结构	段型数目	段型比	段次比	具体例子
(0,2)(1,2)	77861	33.29%	49.57%	及其他，于今年，把风车，办法则，放风筝，同行业，开发出，转化为
(0,2)(1,2)(2,2)	94315	40.34%	34.77%	中华人民，工商行政，研制成功，主要领导，产品质量，今天下午，国家规定
(0,3)(2,2)	6876	2.94%	3.47%	自行车厂，旅游业务，消防队员，房地产业，发电机组，合格证书，政治局面
(0,2)(1,2)(2,3)	5981	2.56%	1.53%	文化工作者，解放生产力，上报国务院，进行规范化，参加座谈会
(0,2)(1,3)	4819	2.06%	1.51%	国外交部，种子公司，促进出口，上天安门，报国务院，美国务卿，和解放军
(0,2)(1,2)(2,2)(3,2)	11308	4.83%	1.37%	国内外贸易，为主要目标，落实在行动，展现在世人，工作主要是
(0,3)(2,2)(3,2)	6784	2.90%	1.34%	共和国土地，联合国难民，标准化工作，大部分地区，进一步调整
(0,2)(1,2)(2,4)	3840	1.64%	0.88%	适应市场经济，严重刑事犯罪，中国有色金属电子集团公司
(0,2)(1,2)(2,2)(3,2)(4,2)	6225	2.66%	0.75%	提高产品质量，对外加工装配，集体统一经营，专职工作人员，内部分配制度

（续表）

(0,4)(3,2)(4,2)	1859	0.79%	0.68%	社会主义理论，古典文学名著，出租汽车行业，经济作物种植，举足轻重地位
(0,4)(3,2)	1420	0.61%	0.60%	前所未有的，高尔夫球场，乡镇企业已，别开生面的，社会保险局
(0,4)(2,4)	182	0.08%	0.59%	市场经济体制，企业集团公司，自成一家之言，星火燎原之势，风云变换莫测
其他	12382	5.30%	2.75%	

(2)微结构类型的描写和分类统计结果

最大交集字段的微结构是指对字段所包含的所有词（不管它们是否为交集因子）的位置和词长（包括词长为1的单字词）。例如：



最大交集字段：重 工 业 区

微结构类型：(0,1)(0,3)(1,1)(1,2)(1,3)(2,1)(3,1)



最大交集字段：棉 花 生 产

微结构类型：(0,1)(0,3)(1,1)(1,2)(2,1)(2,2)(3,1)

一个宏结构通常包含多个微结构。以最简单、最高频的宏结构"(0,2)(1,2)"来说，也有以下8种对应的微结构：

1(0,2)(1,2)　　　"浩浩森"

2(0,1)(0,2)(1,2)　　　"受贿赂"

3(0,2)(1,1)(1,2)　　“榕树桠”

4(0,2)(1,2)(2,1)　　“足协从”

5(0,1)(0,2)(1,1)(1,2)　　“不过敏”

6(0,1)(0,2)(1,2)(2,1)　　“安徽调”

7(0,2)(1,1)(1,2)(2,1)　　“焕发出”

8(0,1)(0,2)(1,1)(1,2)(2,1)　　“水平和”

对 Rcorpus 全部 233888 条最大交集字段按微结构分类，共得到 1667 类。表 5－9 列出了其中最主要的微结构类型。平均每一种微结构类型含约 140 条最大交集字段个例，最高频的两种微结构分别是“(0,1)(0,2)(1,1)(1,2)(2,1)”和“(0,1)(0,2)(1,1)(1,2)(2,1)(2,2)(3,1)”，它们累计的段型比达 71.37%，段次比达 82.35%。统计结果表明，最大交集字段的微结构分布虽比宏结构分布显得更分散，但整体来讲还是相当集中的。

表 5－9　最大交集字段微结构的分布

最大交集字段的微结构	段型数目	段型比	段次比	具体例子
(0,1)(0,2)(0,3)(1,1)(2,1)(2,2)(3,1)	4310	1.84%	2.14%	外交部长，交响乐团，青年人才，生活费用，所在地区，解放军队，时装表演，无线电厂，消防队员，受灾面积

(续表)

(1,2)(2,1)(1,1)(1,2)(2,1)(2,2)(3,1)(3,2)(4,1)	10988	4.70%	1.33%	大会堂会见,百分之一点,充分说明了,开发生产出,地表现出来,近年来由于,生活水平和,于今年年底
(0,1)(0,2)(1,1)(1,2)(2,1)(2,2)(2,3)(3,1)(4,1)	3535	1.51%	1.02%	解放生产力,革命根据地,服装设计师,出生性别比,养老保险费,极端重要性,管理科学化,发展现代化
(0,1)(0,2)(0,3)(1,1)(1,2)(2,1)(2,2)(3,1)	1236	0.53%	0.94%	自行车厂,流窜犯罪,一方面对,输电线路,小学校长,地下水位,带头人和,推动力量,安全部门,面积分别
(0,1)(0,2)(0,3)(1,2)(2,1)(3,1)(3,2)(4,1)	3692	1.58%	0.78%	代表团团长,解放军官兵,成年人犯罪,平方米面积,现代化装备,地下党组织,所有制成分,共产党内部
(0,1)(0,2)(1,1)(1,2)(2,1)(2,2)(3,1)(3,2)(4,1)(4,2)(5,1)	5967	2.55%	0.73%	随着生活水平,开放大米市场,极大地方便了,没有形成规模,外汇收入超过,内部分配方式,总结交流经验,这个中心服务
(0,1)(0,2)(1,1)(1,2)(1,3)(2,1)(3,1)	2048	0.88%	0.62%	在座谈会,有生命力,新生长点,负有心人,和解放军,对开发区,了当事人,还有赖于,本科学家,一代表团

（续表）

(0,1)(0,2)(1,1)(1,2)(2,1)(2,2)(2,4)(3,1)(4,1)(4,2)(5,1)	1999	0.85%	0.57%	中国运载火箭，现有生产能力，紧急电话会议，中国外汇制度，防止水土流失，生产假冒伪劣，严重水土流失，更加深入人心，
(0,1)(0,2)(0,4)(1,1)(2,1)(2,2)(3,1)(4,1)(4,2)(5,1)	111	0.05%	0.57%	成人教育中心，出乎意料之外，拳头产品开发，登山运动健将，不过如此而已，增产增收节支，有限广播电台，瓢泼大雨倾盆，四面八方支援
(0,1)(0,2)(0,4)(1,1)(2,1)(2,2)(3,1)(3,2)(4,1)(4,2)(5,1)	1138	0.49%	0.6%	投入产出水平，自力更生发展，独立自主和平，水土流失重点，大案要案情况，主观能动作用，广播电台联合

文献[96]还采用128个词类标记，分别对最大交集字段的宏结构和微结构进行再分类。例如，最大交集字段“办法则”、“放风筝”和“转化为”的宏结构类型均为“(0,2)(1,2)”，但它们所含交集因子的词性不同，依次为(名词、名词)，(动词、名词)和(动词、动词)。

对Rcorpus中全部233888条最大交集字段的交集因子指派相应的词类标记，然后再按宏结构分类，共得到16548类。表5-10给出了其中12种主要类型的分布情况。按交集因子的词类信息对最大交集字段宏结构作进一步的细分

类，得到的分布是比较分散的，平均每类只含约 14 条个例，最高频的前 12 种宏结构的累计段型比和段次比均未超过 50%。

同理，可以对最大交集字段中的所有词（包括单字词）都标上相应的词类标记，然后再按微结构进行分类。例如，最大交集字段“立法权”对应的微结构和相关词类信息为“(0,1)(0,2)(1,1)(1,2)(2,1)+(vg,vg,ng,ng,ng)”。结果 233888 条最大交集字段共得到 123356 类。平均每类所含的最大交集字段不足 2 条。由于用这种方法对最大交集字段微结构进一步细分所得类别数太大，这种分类的实际意义不大。但这个统计结果也反映出最大交集字段的结构模式是极其复杂的，在具体设计交集段的切分算法时，这个因素仍然是必须考虑的。

表 5-10　最大交集字段宏结构+词类信息的分布

最大交集字段的宏结构	词类	段型数目	静态频率	动态频率	具体例子
(0,2)(1,2)	(ng,ng)	21929	9.38%	12.79%	族人民，深层次，工人们，年历史，一时期，假发票
(0,2)(1,2)	(vg,vg)	9644	4.12%	5.68%	转化为，不适应，适用于，和解决，奉献给，应承担

（续表）

(0,2)(1,2)(2,2)	(ng,ng,ng)	12579	5.38%	5.32%	中华人民，产品质量，汽车工业，干部队伍，内科学会
(0,2)(1,2)	(vg,ng)	11971	5.12%	4.49%	出国门，等同志，主战场，上台阶，从政治，着眼泪
(0,3)(2,2)	(ng,ng)	4602	1.97%	2.80%	基督教徒，价值观念，交响乐团，摩托车队，坐标系统
(0,1)(1,2)(2,2)	(vg,ng,ng)	7578	3.24%	2.55%	执行情况，施工人员，跻身世界，实行市场，侵犯人权
(0,2)(1,2)(2,2)	(vg,vg,ng)	9123	3.90%	2.53%	与会同志，施加压力，实现形式，筹集资金，发生事故
(0,2)(1,2)	(ng,vg)	5745	2.46%	1.90%	需求和，地支持，人参加，地理解，文化为，学校对
(0,2)(1,2)(2,2)	(s,ng,ng)	1761	0.75%	1.45%	中国文化，长江流域，香港客人，上海市场，西藏历史

（续表）

(0,2)(1,2)(2,2)	(vg,vg,vg)	4917	2.10%	1.39%	联合办学，公开拍卖，引起重视，检查验收，装修改造
(0,2)(1,2)	(vg,a)	970	0.41%	1.39%	起重要，高清楚，打破旧，和亲切，提高大，分神秘
(0,2)(1,2)	(wg,a)	965	0.41%	1.35%	越发展，不断档，已经办，着实现，的确立，向来访
其他		142104	60.76%	56.36%	

综上所述，自动分词系统对交集型歧义切分字段的处理效果迄今不尽人意，其主要原因是对交集字段的复杂性胸中无数，因而采用的排歧算法都过于粗糙。山西大学和清华大学在大规模语料库上对汉语交集字段所做的普查工作，用翔实的统计数据揭开了交集字段的神秘面纱，让人们既看到这类处理对象的共性，又触摸到它们形态结构各异的个性。这对于有针对性的设计自动排歧的分词算法是十分必要的。这两个研究小组的工作表明，在交集字段语料库调查的基础上设计的自动分词系统，对歧义切分字段的处理精度得到了实质性的提高。

第二节　汉语基本名词短语识别研究

基本名词短语的识别是自然语言处理、信息检索和机器翻译等领域的一项基础研究。丘奇(Church)把英语的基本名词短语(baseNP)定义为“非嵌套的名词短语”[97]，把 baseNP 的识别看作是对其左右边界的标注问题，并利用 N 元模型来实现。香港中文大学李文捷[98]曾用 N 元模型对汉语的最长名词短语进行识别。她们的实验表明，仅仅依赖词类信息的 N 元模型不足以正确识别汉语文本中的名词短语。清华大学赵军[99]定义了汉语的基本名词短语，并从人工标注的训练语料库中抽取出 baseNP 的上下文无关句法组成模板。但是研究表明，符合句法组成模板的词语序列只是构成 baseNP 的必要条件，而不是充分条件。如果将文本中符合基本模板的词语序列全部判定为 baseNP，正确率只有48.5%。因此，为了正确识别文本中的 baseNP 还需要通过训练获取 baseNP 的上下文有关规则，并据此实现 baseNP 的自动识别系统。研究表明，把 baseNP 的基本组成模板与表示 baseNP 上下文环境的转换规则结合起来，在封闭测试和开放测试两种情况下，baseNP 的识别正确率分别达到了 91.1% 和 87.3%。

一、汉语 baseNP 的定义

如前所述，丘奇(Church)把英语基本名词短语定义为"非嵌套的名词短语"，即它的内部不再包含更小的名词短语。这一定义显然不能满足中文信息处理的要求，例如："自然语言处理"、"亚洲金融危机"和"经济体制改革"等显然都不是"非嵌套的名词短语"，但对于信息检索和机器翻译来说都是具有特定意义的、需整体看待的名词短语。张卫国曾把汉语名词短语的定语分为三种类型：限定性定语、描写性定语和区别性定语[100]。赵军遵照限定性定语的概念提出了汉语基本名词短语(baseNP)的如下形式化定义：

BaseNP → baseNP + baseNP

BaseNP → baseNP + 名词|名动词

BaseNP → 限定性定语 + baseNP

BaseNP → 限定性定语 + 名词|名动词

限定性定语 → 形容词|区别词|动词|名词|处所词|西文字串|数量词

根据上述定义，汉语名词短语可分为 baseNP 和 ~baseNP(非基本名词短语)两大类，以下举例说明。

表 5-11 BaseNP 和～baseNP 示例

BaseNP	～baseNP
甲级联赛 产品结构 空中走廊 下岗女工 促销手段 太空旅行 自然语言处理 企业承包合同 第四次中东战争	复杂的特征 这台计算机 很大成就 对于形势的估计 明朝的古董 11 万职工 高速发展的经济 研究与发展 老师写的评语

二、baseNP 的句法组成模板

从 baseNP 的定义可知，baseNP 应当遵守一定的句法规则，赵军把这种建立在词类及短语标记基础上的上下文无关规则叫做句法组成模板(简称模板)。但是，进一步的观察表明，符合句法组成模板的词语序列只是构成 baseNP 的必要条件，而不是充分必要条件。一个符合模板的词语序列可以不是 baseNP，这里有以下两种情况：

1. 边界模糊：在一个句子中，某些符合模板的词语序列可能是语法形式，也可能是非语法形式。请考察如下两个例句：

例 1：技术 改造 是 国营 企业 走出 困境 的 出路。(baseNP)

例 2：IBM 公司 宣布 全面 降低 个人 电脑 的 销售 价格。(非语法形式)

例 1 中的"技术/N 改造/V"和例 2 中的"公司/N 宣布/V"都符合下列 baseNP 模板：BaseNP→N + V。但前者是

baseNP,后者不仅不是 baseNP,而且是非语法形式。换言之,“公司”和“宣布”这两个相邻词在例 2 中分别属于主语和谓语,它们不在一个短语结构内。

2. 短语类型不同:在一个句子中,某些符合模板的词语序列虽然是语法形式,但可能是 baseNP,也可能是其他类型的短语。在下面两个例句中:

例 3:今年/T 大学/N 毕业生/N 的/U 就业/V 形势/N 严峻/A 。

例 4:中国/N 人民/N 银行/N 今天/T 宣布/V 降低/V 利率/N 。

“就业/V 形势/N”和“降低/V 利率/N”这两个词语序列都符合 baseNP 的另一个模板:BaseNP → V + N。但前者是 baseNP,后者却是一个动词短语。

文献[99]介绍了赵军进行汉语 baseNP 识别的两步策略:(i)从人工标注了 baseNP 的训练语料中抽取 baseNP 的上下文无关句法组合模板,把测试文本中符合这些模板的词语序列作为候选的 baseNP;(ii)利用转换学习方法从语料库中获取 baseNP 的上下文有关规则,据此判断哪些候选的词语序列是真正的 baseNP。

三、baseNP 句法组成模板的抽取

模板的抽取工作分两步进行:1. 建立人工标注 baseNP

的语料库;2. 根据语料库的统计信息,对初始模板集合进行筛选,形成基本组成模板集合。

1. baseNP 语料库的标注

赵军人工标注 baseNP 的语料库为 10 万字次,在对它实行自动分词和词性标注的基础上,根据 baseNP 的定义和以下几条规范,进行 baseNP 的人工标注。这些规范进一步明确在 baseNP 中不包含结构助词"的"、表示并列关系的连词(如"和"、"与"、"及"、"以及"等)和顿号、时间词、代词、介词以及基数词和量词的组合等。

2. baseNP 的基本模板

在词语的词性和音节信息的基础上,赵军从人工标注 baseNP 的训练语料库中统计得到 407 个 baseNP 句法组成模板,其中出现次数超过 5 次的有 64 个,它们覆盖了语料库中 98.6%的 baseNP,赵军把他们称为基本模板。下表 5－12 列出一些常用的基本模板,模板中的每个句法标记由两部分组成:前面的英文大写字母串表示词语的词性(详见附录),后跟的数字表示词语的音节数。例如:模板"NG2 + VN2"表示由双音节普通名词(NG2)和双音节名动词(VN2)组成的 baseNP。

表 5－12　baseNP 的基本组成模板例表

模板	示例	模板	示例
VN2 + NG2	教育理论 调查报告	NG2 + NG2 + VN2	产品结构调整 住房制度改革

（续表）

VGN2 + NG2	委托方式 打击力度	NG2 + VN2 + NG2	情报检索方法 概率标引模型
VG02 + NG2	下岗女工 促销手段	NG2 + NG2 + NG2	脂肪分子结构 热带草原气候
S2 + NG2	空中走廊 海底光缆	NG2 + NG2 + VN2 + NG2	政治体制改革进程 思想品德教育理论
XCH + NG2	0157 病毒 IBM 公司	NG2 + NG2 + VG02 + NG2	经济信息流通方式 航天飞机飞行计划

统计显示，如果将文本中所有符合基本模板的词语序列全部标注为 baseNP，召回率是 98.6%，而精确率仅为48.5%。这说明仅仅依靠上下文无关模板不能解决 baseNP 识别中边界模糊和短语类型这两类歧义。

四、识别 baseNP 的上下文有关规则

布里尔(Brill)提出一种由错误驱动的转换学习方法[101]，拉姆肖(Ramshaw)曾将这种方法应用于英语文本的短语边界识别[102]。赵军利用布里尔的转换学习方法来获取识别 baseNP 的上下文有关规则。其算法的流程图如图 5－3 所示。

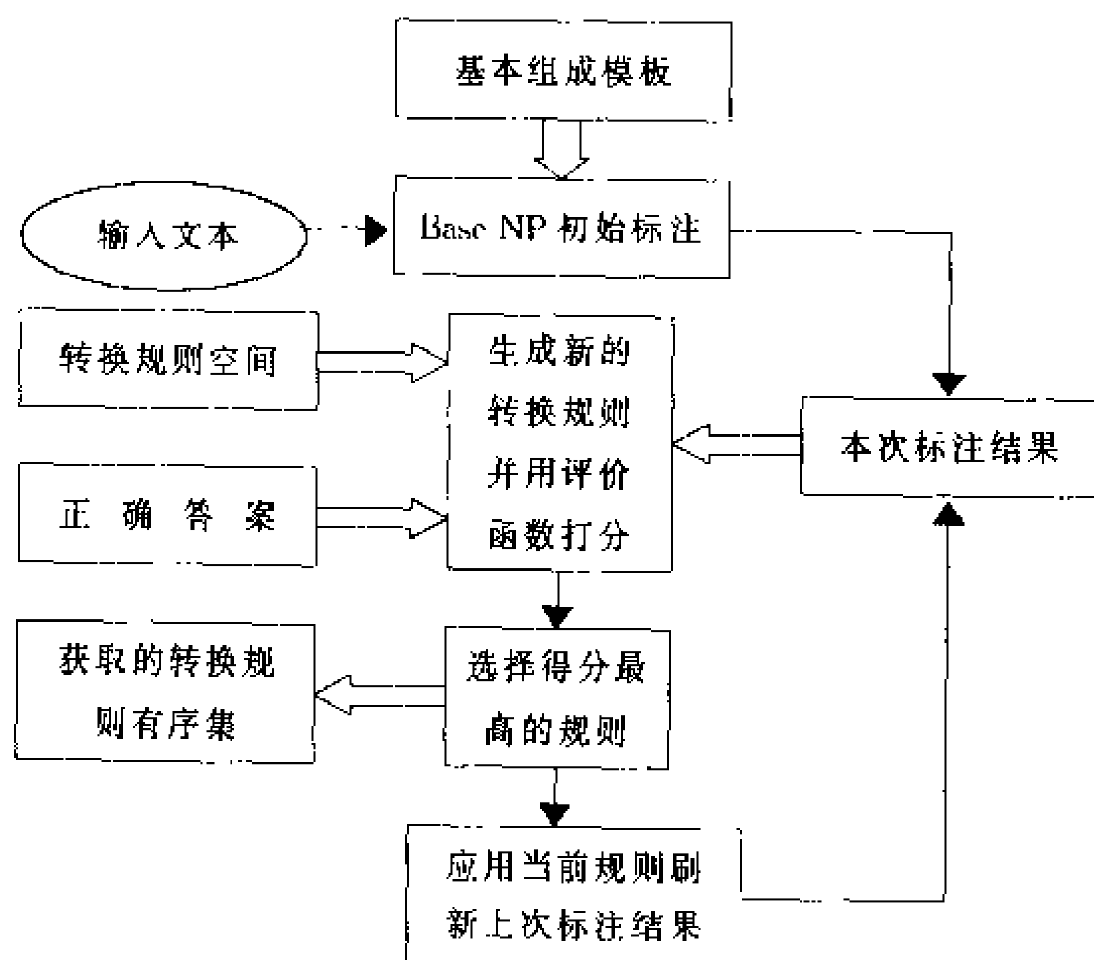


图 5-3 基于转换的 baseNP 识别模型示意图

首先根据基本模板对输入文本进行 baseNP 的初始标注，然后把初始标注结果与正确答案加以比较，以发现标注中的错误。查阅事先定义的上下文有关转换规则的模板，从中挑选出能纠正当前标注错误而且得分最高的一条规则。用这条规则对当前标注结果整个进行转换，并把这条新获取的规则顺序地存放于规则集中。循环地用上述方法从错误标注中学习到一条条新的上下文有关的转换规则，形成一个有序的转换规则集。综上所述，这种转换学习方法有三个关键模块：

1. 初始标注模块

利用基本模板对输入文本进行 baseNP 的初始标注，得到

baseNP 的候选集。标注过程如下：

假设输入文本中每个词记作 w_i，其词性记作 t_i，则输入文本可表示为如下的一个符号串：

$$w_1/t_1\ w_2/t_2 \ldots w_i/t_i \ldots w_j/t_j \ldots w_{N-1}/t_{N-1}\ w_N/t_N$$

如果在基本模板中存在这样一条上下文无关规则：

$$t_i \ldots t_j, \longrightarrow baseNP,$$

则将其中的子串 $w_i/t_i \ldots w_j/t_j$ 初始标注为 baseNP。例：

美国/NF学者/NG 提出/VGN 一/MX 种/QN

概率/NG 标引/VN 方法/NG。/。

在上面这个输入句子中，用实线分别标出了 6 个 baseNP 候选词串。

2. 转换规则模板

每一条转换规则模板包括转换动作和触发环境两个要素，其中触发条件规定了上下文环境的若干特征，转换动作则对文本中原有的某个标注结果进行更新。具体来说，转换动作就是在 baseNP 的候选标记、确认标记和否定标记三者之间的转换，如：

(1) 转换动作 1：将词串 W 上的候选标记转换为确认标记；

(2) 转换动作 2：将词串 W 上的候选标记转换为否定标记；

(3) 转换动作3:将词串 W 上的确认标记转换为否定标记;

(4) 转换动作4:将词串 W 上的否定标记转换为确认标记;

转换规则的触发条件规定为 baseNP 候选词串的前两个词和后一个词,以及该候选词串中的第一个词和最后一个词。对构成触发条件的这些相邻词,要考察它们的词性、义类和音节数等特征,同时也要兼顾当前候选词串属于哪种模板类型。具体来说,共有以下20种不同的触发条件:

(1) $POS(p_{-1})=t$;

(2) $POS(p_1)=t$;

(3) $POS(p_{-2})=t$;

(4) $SENSE(p_{-1})=s$;

(5) $SENSE(p_1)=s$;

(6) $SENSE(p_{-2})=s$;

(7) $SYL(p_{-1})=x$;

(8) $POS(p_{-1})=t_1$.AND.$POS(p_1)=t_2$;

(9) $POS(p_{-2})=t_1$.AND.$POS(p_{-1})=t_2$;

(10) $POS(p_{-1})=t_1$.AND.$SENSE(p_{-1})=s_1$

(11) $SP(W)=m$.AND.$POS(p_{-1})=t$;

(12) $SP(W)=m$.AND.$POS(p_1)=t$;

(13) $SP(W)=m$.AND.$POS(p_2)=t$;

(14) $POS(p_{-1})=t_1$.AND.$POS(BEGIN(W))=t_2$;

(15) $POS(p_1)=t_1$.AND.$POS(END(W))=t_2$;

(16) $SENSE(p_{-1})=s_1$.AND.$SENSE(BEGIN(W))=s_2$;

(17) $POS(p_{-1}) = t$ .AND. $SENSE(p_{-2}) = s_1$ .AND. $SENSE(BEGIN(W)) = s_2$;

(18) $SENSE(p_{-1}) = s_1$ .AND. $SENSE(END(W)) = s_2$;

(19) $POS(p_{-1}) = t$ .AND. $SENSE(p_{-2}) = s_1$ .AND. $SENSE(END(W)) = s_2$;

(20) $SENSE(p_1) = s_1$ .AND. $SENSE(END(W)) = s_2$;

其中 W 表示当前的候选词串，p_{-2}，p_{-1}，p_1 分别表示 W 的前二、前一和后一位置上的词；$POS(p)$、$SENSE(p)$和 $SYL(p)$分别表示位置 p 的词的词性、义类和音节数，义类代码取自《同义词词林》的大类、中类和小类。$SP(W)$表示候选词串 W 所属的模板类型，$BEGIN(W)$和 $END(W)$分别表示 W 的第一个词和最后一个词。此外，如果处于位置 p 的词已经属于一个 baseNP，则记作 $POS(p)$ = BN。

请注意，在以上列出的不同触发条件中，t、x 和 s 是分别表示词性、音节数和义类代码的变量，它们的值要通过训练文本来确定。因此，在这些变量的值未被确定之前，转换规则还仅仅是一个尚待变量赋值的模板。这些转换规则模板的集合定义了全部转换规则的可能形式，叫做转换规则空间。转换学习的目的就是通过人工标注的语料来给规则模板中的变量赋值，使它们变成程序可执行的转换规则。所以在这里区分转换规则模板和转换规则这两个概念是必要的。

3. 评价函数

为了通过训练语料选出标注结果最好的转换规则，需要

定义一个评价函数来给每条转换规则打分。假设对当前文本应用任意一条转换规则 r，如果应用后所得新文本的标注正确率最高，便判定它的得分最高，并且令它成为这一轮转换中的转换规则。换言之，设转换规则 r 把当前文本中的错误标记更正为正确标记的次数为 $C(r)$，同时使当前文本的正确标记误改为错误标记的次数为 $E(r)$，则评价函数给该规则的评分为：

$$F(r)=C(r)-E(r)$$

五、转换规则的学习算法

学习算法的目的是从事先定义的转换规则空间中自动生成一组有序的上下文相关的转换规则。在算法的每次循环中，学习器将遍历所有满足触发条件的转换规则模板，根据应用一条转换规则刷新当前文本后所得的标注结果，用评价函数为它打分，把得分最高的转换规则作为本次循环的结果，并把它顺序装入有序的规则集中（第一次循环获得的规则位于规则集的首位，第二次循环获得的规则位列次位，等等），接着用这条新的转换规则刷新当前文本的相关标记，得到一个新的当前文本，进入下一轮循环。学习过程如此循环进行，直到再找不到一条转换规则能使其得分高于某个阈值，学习过程终止。如上所述，学习过程获得的转换规则集是有序排列的，前一次选中的规则总是位于后一次选中的规则的前面。对任

意一个输入文本进行 baseNP 标注时，也先用初始标注模块对输入文本进行初始标注，然后依次使用转换规则集合中的转换规则，对当前文本进行标注的刷新。

转换规则的学习算法可阐述如下：

设 C 是未标注 baseNP 的语料库，C_A 是对 C 进行了人工标注 baseNP 的语料库，TS 是有序的转换规则集，起始时令 $TS=\varphi$（空）。

a.应用 baseNP 的基本句法组成模板对 C 进行初始标注，得到当前标注文本 $C^{[0]}$；

b.重复以下步骤，直到不再存在转换规则 r，能使 $F(r)>T$（T 为得分阈值）。

第 i 次循环（$i=0,1,2,\cdots$）

1）对比 $C^{[i]}$ 和 C_A，找出 $C^{[i]}$ 中错误标注集合 $E^{[i]}$；

2）以 $E^{[i]}$ 驱动，在转换空间中搜索一条最佳的转换规则 $r^{[i]}$，使 $r^{[i]}$ 的评分最高，

$$r^{[i]}=\underset{r}{\mathrm{argmax}}\mathrm{F}(r);$$

3）将 $r^{[i]}$ 添加到 TS 的最后，并用 $r^{[i]}$ 刷新 $C^{[i-1]}$，得到 $C^{[i]}$。

六、实验结果

实验分以下三部分：

(1)从训练语料库中抽取 baseNP 的基本句法组成模板；

(2)利用错误驱动的学习算法从训练语料库中获取 baseNP 的上下文有关转换规则；

(3)将基本组成模板与上下文有关的转换规则结合起来识别文本中的 baseNP。

其中有关第(1)部分的实验结果在本节(5.3)中介绍，下面分别介绍第(2)(3)部分的实验结果。

a. 上下文有关转换规则的获取实验

用错误驱动的转换学习方法，从 10 万字的训练语料库中，令评价函数的阈值 $T=0$，总共获得 380 条上下文有关的转换规则。下面列出其中 10 条最常用的转换规则：

1) *change* 候选标记 *to* 确定标记

when $POS\ (p_{-1})=QN$.AND. $POS\ (p_1)=$ 。

例：该/R 公司/NG 今年/T 与/CM 外商/NN 签定/VN 两/A 项/QN [承包/VNN 合同/NG]。/。

2) *change* 候选标记 *to* 确定标记

when $POS(p_{-1})=CM$.AND. $POS\ (p_{-2})=BN$

例：……将/P 各/R 种/QN[反/H 坦克/NG 火力/NG]和/CM[防/H 坦克/NG 障碍物/NG]密切/A 结合/VGN……

3) *change* 候选标记 *to* 确定标记

when $POS\ (p_{-1})=$“ .AND. $POS\ (p_1)=$”

例：这/R 种/QN 语法/NG 已经/D 成为/VGN 许多/MG 立足/VGO 于/P “/“[复杂/A 特征/NG]” /” 的/USDE “/“[合一/NG 运算/VNN]”/” 的/USDE[形式化/VNO 方法/

NN]的/USDE 基础/NG 。/。

4）*change* 候选标记 *to* 确定标记

when *SENSE*(p_{-1}) = Ja02

例：这/R 种/QN 气候/NG 叫做/VGN/Ja02[热带/NG 雨林/NG 气候/NG]……

5）*change* 候选标记 *to* 确定标记

when *POS* (p_{-1}) = *P* .AND. *POS* (p_1) = *V*

例：在/P[上海/NG 战役/NG]结束/VGO 后/F ，……

6）*change* 候选标记 *to* 确定标记

when *POS* (p_{-1}) = M

例：许多/MG[亏损/VGO 企业/NG]将/D 转产/VGO ，……

7）*change* 确定标记 *to* 否定标记

when *POS* (p_{-1}) = M .AND. *POS* (p_1) = U

例：许多/MG 地方/NG 分布/VN 着/UT 茂密/A 的/USDE 热带/NG 雨林/NG ，……

8）*change* 候选标记 *to* 否定标记

when *POS* (p_{-1}) = *D* .AND. *POS*(*BEGIN*(*W*)) = VGN

例：两/MJ 国/NG 政府/NG 今天/T 联合/D 发表/VGN/Hc11 建交/VNO 公报/NG/Dk14，……

9）*change* 候选标记 *to* 确定标记

when SENSE (p_{-1}) = Hc11 .AND. SENSE(*END*(*W*)) = Dk14

例：两/MJ 国/NG 政府/NG 今天/T 发表/VGN[建交/

VGO 公报/NG],……

10) *change* 候选标记 *to* 确定标记

when SENSE(p_{-1}) = *Ie*02 .AND. *POS* (p_1) = 。

例:… 组成/VGN/ Ie02[防/H 步兵/NG 火力/NG 配系/NG]。/。

从上述示例中可以看到,错误驱动的学习方法所获取的转换规则是确定性的。如果孤立地看某一条规则,它可能不完全正确(如:规则示例6),即规则的转换动作并不完全适用于其触发环境。但是每条规则的使用都是确定性的,由它引起的少量错误标记可以通过后续的转换规则(如:规则示例7)来弥补。因此,整个规则集是有序的,排列在前面的规则更具归纳性,而后续的规则则更有特殊性。

b. baseNP 识别的实验

baseNP 识别系统的工作步骤如下:

1)依据 baseNP 的基本句法组成模板对输入文本进行 baseNP 的初始标注;

2)依次使用转换规则集中的转换规则对输入文本的标注结果进行转换;

3)在最终的标注结果中,如果仍然存在歧义标注,即某个词语序列对应两个或两个以上的确定标记,如:

词串: $w_1/ t_1\ w_2/ t_2 \ldots w_i/ t_i \ldots w_j/t_j \ldots w_{N-1}/t_{N-1}\ w_N/t_N$

标记1: ————

标记2: ——————

则只保留其中最长的序列所对应的标记，删除其余标记。

为了对 baseNP 的识别进行对比研究，赵军把 baseNP 的实验分成两组。第一组实验只应用 baseNP 的基本句法组成模板，这组实验的结果可以看作是 baseNP 识别的基线（baseline）。第一组实验顺序执行上述实验过程的第 1)、3)步。第二组实验采用基于句法组合模板与上下文有关转换规则相结合的方法，即顺序执行上述实验的第 1)、2)、3)步。

两组实验都分成封闭测试和开放测试两部分，封闭测试和开放测试的输入文本规模都为 10000 字次，其中封闭测试文本选自训练语料库，开放测试文本来自训练语料库之外的语料。

测试 baseNP 识别系统的性能指标规定为精确率 P 和召回率 R：

$$\text{精确率}: P = \frac{a}{b} \times 100\%$$

$$\text{召回率}: R = \frac{a}{c} \times 100\%$$

其中，a 是输入文本中被系统正确识别的 baseNP 数，b 是被系统判定为 baseNP 的词语序列总数，c 是输入文本中实际存在的 baseNP 数。

表 5－13 是两组实验的评测结果：

表 5-13　两种 baseNP 识别方法的对比

测试 类型	只用基本组成模板的方法		基本组成模板与转换规则结合的方法	
	精确率	召回率	精确率	召回率
封闭测试	72.9%	77.5%	91.1%	93.2%
开放测试	72.7%	78.3%	87.3%	91.8%

从上述两组实验结果的对比中可以看出，基本组成模板与上下文有关转换规则相结合的 baseNP 识别方法明显优于单纯基于组成模板的 baseNP 识别方法。

以下给出了开放测试中一个输入文本的部分标注结果(加黑的部分为错误标注)。

[干部/NG 工作/NG]是/VY 中国人民解放军/NG 依据/P 中国共产党/NG 的/USDE[干部/NG 路线/NG]和/CM[政策/**NG 管理/VNN 军官/NG**]和/CM[文职/NG 干部/NG]的/USDE 工作/Ng。/。[机构/NG 干部/NG 工作/NG],/,原来/D 是/VY[中国人民解放军/NG 建设/VNN]的/USDE[重要/A 内容/NG]o/o 根据/P1929 年/T[古田/NPL 会议/NG 决议/NG]的/USDE 规定/NG,/,[工农/NG 红军/NG]的/USDE[军事/NG 干部/NG]由/P[军事/NG 系统/NG]管理/VNN,/,其/R[具体/A 工作/NG],/,由/P[司令/NG 机关/NG]的/USDE[队列/NG 部门/NG]和/CM[政治/NG 机关/NG]的/USDE[组织/NG 部门/NG]负责/VGV,/,1973 年/T 以后/F,/,干部/NG 的/USDE 任免/VNN、/、调配/VNN 由/P 各/R 级/NG[军政/NG 委员会/NG]按/P[任免/

VNN 期限/NG]讨论/VNN 决定 VGN,/,有的/R 部队/NG 还/D 在/P 队列/NG 和/CM[组织/NG 部门/NG]内/F 成立/VGN 了/UT 干部科/NG 。/o

第三节 基于结构词义空间的汉语词义排歧模型

词义排歧(word sense disambiguation)指根据一个多义词在文本中出现的上下文环境来确定其词义代码。这个代码可以是该词义在一部普通词典中的义项号,也可以是它在一部义类词典中的义类代码,还可以是翻译词典中一个词所对应的目标词或概念词典中词的定义项。长期以来,词义排歧一直被认为是自然语言处理的一个难题。90年代以前,词义排歧研究主要采用人工智能方法,其困难在于要用人工来编制大量的排歧规则,不仅覆盖面很窄,而且开销巨大,即所谓知识获取的"瓶颈"问题。90年代以后,由于大规模机器可读词典和语料库的出现,词义排歧研究进入了一个以语料库方法为主的新时期。

基于机器可读词典的词义排歧方法充分利用普通词典中词条的释义文本,通过计算一个多义词各义项的释义文本与当前文本的重叠程度来实现排歧目的,如莱斯克(Lesk)和威尔克斯(Wilks)分别提出的词义排歧方法[103,104]。但实际上

当词条的释义文本比较短时，比如只用近义词或反义词来释义，则在该词出现的当前文本中很难找到与释义文本重叠的信息，从而影响了词义排歧的效果。另外一类基于词典的词义排歧方法是利用义类词典。其中具有代表性的工作是亚罗斯基(Yarowsky)提出的词义排歧算法[105]。这种方法在计算每个语义类的凸显词(salient words)时将多义词的每次出现平均地分到多义词所对应的各个语义类中，从而引入统计噪声；此外寻找凸显词所使用的语料范围受限，因而覆盖面较窄。

基于语料库的词义排歧方法(Yarowsky 1994，Bruce 1995等)[106,107]大都需要对训练语料库进行词义的人工标注，这种标注工作既费时又费力，并且统计结果存在严重的数据稀疏问题。因此一部分学者致力于研究无指导的(unsupervised)词义排歧知识获取方法。但迄今这些方法大都只停留在几个或十几个多义词的小规模实验上。清华大学李涓子提出了一种基于结构词义空间的词义排歧模型。由于在《同义词词林》中每个同义词集对应于一个唯一的义类代码，而在一个同义词集中，一般来说，既有少量多义词，又有大量单义词。因此可以在一个大规模语料库上统计并抽取任一同义词集中那些单义词前后同现的实词来构造这个义类代码的“分类器”。由于这种词义排歧知识的学习是无指导的，从而可以免除对语料库实行词义人工标注的大量开支。实验表明，这种词义排歧模型具有较高的排歧正确率，而且对不同领域的

文本具有较好的可移植性。下面介绍她的研究工作。

一、《同义词词林》简介

《同义词词林》[108](简称《词林》)的编者在确定词的语义分类时,以词义为主,兼顾词类,并充分注意题材的相对集中。这部义类词典把词义分为大、中、小类三级,共得到 12 个大类,94 年中类,1428 个小类;小类以下再按同义词词群设立标题词,共含 3925 个标题词。

《词林》中用第一个大写英文字母作为大类的编号,紧接着用第二个小写英文字母表示中类,义类代码的第三和第四位是两个阿拉伯数字,用来表示小类的编号。小类以下的标题词还可细分为两个层次,各用两位阿拉伯数字表示。例如,词"觉悟"的义类代码为"Ga15",其中大类编码 G 表示"心理活动",中类编码 Ga 表示"心理状态",小类编码是 Ga15,它在《词林》中的内容显示为:

Ga15 醒悟 懂事

醒悟 觉悟 省悟 憬悟 觉醒 清醒 如梦初醒 大梦初醒……

懂事 记事儿 开窍 通窍

也就是说,Ga15 包含两个标题词:"醒悟"和"懂事",分别代表这一小类以下的两个词群。因此,词"觉悟"的完整义类代码是 Ga1501.

《词林》以词的义项为收词单位，多义词按其词义被赋以不同的义类代码。例如，词“材料”在《词林》中有三个义项：(1)可以直接造成成品的东西；(2)提供著作的内容的事物或可供参考的事实；(3)比喻适于做某种事情的人才。它们对应的义类代码分别为“Ba06”、“Dk17”、和“Al03”。在文本中对“材料”一词的词义排歧过程，就是根据该词出现的上下文给它标注一个相应的义类代码。

如上所述，《词林》的义类代码系统构成了一幅有层次结构的树状图，如图 5－4 所示。

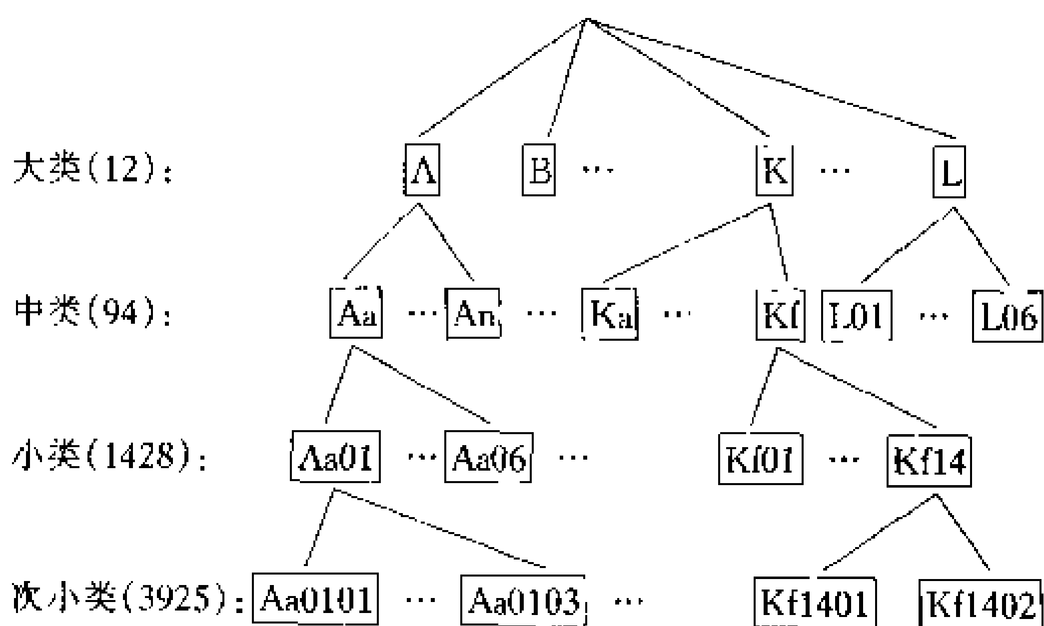


图 5－4：《词林》义类代码系统构成的树状图

《词林》收入的词条实际上包括了词和部分短语、成语、俗语，总共 50154 条。《词林》中多义词的分布情况如表 5－14 所示。统计表明，《词林》中总共有 7430 条多义词，占词条总数的 14.8%，即七分之一强。值得注意的是，仅占词条总数 7.52% 的 3774 条单字词中，却有近一半（即 1801 条）是多义

的；相比之下，在 46380 条多字词中，只有 12.1% 是多义的。

表 5－14　《词林》中多义词的分布情况

	单字词		多字词		词条总数
	词条数	百分比	词条数	百分比	
单义词	1973	52.3%	40751	87.9%	42724
多义词	1801	47.7%	5629	12.1%	7430
总计	3774	100%	46380	100%	50154

词义排歧任务的难易程度还可以通过语料库的调查来显示。一般来说，在一个语料库中，总词次的大约 42% 具有不只一个词义。如前所述，《词林》的语义分类兼顾了词的词性，比如 A－D 四个大类多属名词，数词与量词归入中类 Dn，E 类多属形容词，F－J 类多属动词等。因此，对于一个经过分词和词性标注的输入文本来说，其中为数不少的多义词可以直接根据它们的词性来辨识词义。统计数据显示，经过词性排歧后，语料库中多义词的词次数占总词次数的比例从 42% 降低到 24%，减少幅度可达 43%。

二、《词林》的向量空间表示

"观其伴而知其意"（You shall know a word by the company it keeps）是语言学家弗斯（Firth 1957）[109]对词义辨识的描述。也就是说，一个词的词义只能在它的应用中得以辨识。因此，如果一个词的某个特定词义在一个语料库中多次出现，对其每次出现的上下文加以考察，就可以获得该词

义同其他词的搭配关系。不仅不同的词具有不同的搭配关系，而且同一个词的不同义项也会有不同的搭配关系。

由于一个词的词义可以用与其同现的一组搭配词（简称搭配）来描述，因此在数学上可以用一个多维向量来表示一个特定的词义。李涓子把这样的向量定义为词义向量。具体来说，一个词义向量由多个分量组成，其中每个分量代表与这个词义同现的一个搭配实词，并成为整个词义空间的一维。

考虑到工程实现的需要，不妨把一个词义在文本中出现的“上下文”规定为在这个词前后 d 个位置上同现的全部实词，$\pm d$ 又称为考察搭配的观察窗口。由于这些实词在体现（或辨识）这个特定词义的能力方面不尽相同，因此有必要用一定的权值（weight）来标识它们各自的能力。李涓子把任意一个搭配实词 x_i 与一个特定词义 s 的同现概率 $P(s, x_i)$ 定义为该实词在词义向量中的权值。显然，$P(s, x_i)$ 可以通过语料库的统计来估值。

经过以上修正后，一个词义向量 V 的每个分量就可以用代表这个搭配实词 x_i 的同现概率 $P(s, x_i)$ 来表征，即 $Vx_i = P(s, x_i)$。因此词义向量实质上是一个多维的实值向量，由众多词义向量构成的词义空间便成为一个多维的实值向量空间。

对词义的上述描述，实际上以如下两个基本假设为依据：

［假设 1］ 如果两个词的词义相似，则它们在文本中出现的上下文也相似。若用词义向量分别表示这两个词的上下文，则

它们在词义空间中的距离相近。

[假设 2] 意义相同或相近的一些词，在词义空间上体现为一个密集的点阵。

假设 2 的可靠性可以通过词义的聚类来验证，目的是检验按词义向量描述的词义，通过聚类所得到的同(近)义词集和《词林》的分类体系是否一致。

李涓子的实验是这样设计的。从《词林》中任选两个词类相同的语义小类 A 和 B，设 C_A 和 C_B 分别表示类 A 和 B 中全体单义词组成的集合。即

$$C_A = \{WA_1, WA_2, \cdots, WA_m\}$$

$$C_B = \{WB_1, WB_2, \cdots, WB_n\}$$

其中，$WA_i(i = 1, \cdots, m)$是类 A 的一个单义词，$WB_j(j = 1, \cdots, n)$是类 B 中的一个单义词。依照词义向量的构造原理，可以在一个大规模语料库上分别获得上述任意一个单义词的词义向量 $V(W)$。然后对词表 $C = C_AUC_B$ 中的所有词，按词义向量之间的距离远近重新分类，聚类结果也是两个词集 C_1 和 C_2，有 $C = C_1 \cup C_2$，且 $C_1 \cap C_2 = \Phi$。如果 C_1 和 C_2 分别在一定程度上对应于 C_A 和 C_B，则表明假设 2 是合理的。

聚类时采用自底向上的最短距离算法，首先把词表 C 中在语料库中出现次数大于 100 的词分别归入词集 C_1 和 C_2，然后再依次对低频词进行聚类。

聚类实验采用的语料库规模为 72 兆字节。聚类实验选

用的义类代码对及其在语料库中出现的情况如表 5－15 所示。实验结果如表 5－16 所示。通过词义向量聚类的结果与《词林》编码之间的一致率定义为：

$$\text{一致率}=\frac{\text{聚类结果中与《词林》编码一致的词数}}{\text{词表 C 的总词数}}$$

表 5－15 《词林》的一些义类代码对在语料库中的出现情况

义类代码对	单义词数	总词次数	＞100 次	＞50 次	＞10 次
Hc11/Hc03	17/18	6005/6538	6/7	9/9	13/11
Ba06/Da19	16/16	3415/3954	5/4	6/5	10/8
He11/Hi03	17/18	6005/6165	6/5	9/7	13/13
Aa03/Ae07	15/20	6800/6735	6/4	6/4	10/9
Di10/Di08	27/28	12017/11531	8/8	11/9	20/3
Ed29/Ed11	15/17	3534/4054	2/3	3/6	8/10
Ed16/Ef08	14/17	2599/2656	1/3	2/4	5/8
Gb15/Hj20	7/6	2303/2003	4/2	4/3	6/5

其中，“义类代码对”指从《词林》中选出的一对语义类的名称，“单义词数”是每个义类代码所代表的同义词集中所含单义词的个数，“总词次数”是这些单义词在语料库中的累计出现次数，“＞100 次”表示各个义类代码在语料库中出现次数超过 100 次的那些单义词的个数，其余类推。

表 5－16 聚类结果与《词林》编码的一致率

义类代码对	＞100 次	＞50 次	＞10 次	平均一致率
Ba06/Da19	100％	100％	100％	100％
Aa03/Ae07	90.0％	90.0％	84.4％	82.5％
Di10/Di08	87.5％	85.0％	75.8％	75.5％
Gb15/Hj20	100％	85.8％	81.8％	84.6％

（续表）

Hc11/Hi03	90.9%	82.0%	76.4%	69.3%
Hc11/Hc03	100%	91.7%	91.7%	77.1%
Ed16/Ef08	100%	88.9%	83.3%	81.3%
Ed29/Ef11	100%	100%	84.6%	78.6%
平均一致率	96.1%	90.4%	84.8%	81.1%

实验结果表明：

(1) 根据词义向量之间的距离远近，对语料库中出现次数大于 100 次的单义词进行聚类，有 90% 以上的词与《词林》的结果一致，平均一致率高达 96.1%。对出现次数超过 50 次的单义词，也有 82% 以上的词与《词林》一致，平均一致率为 90.4%。这反映了假设 2 的合理性。

(2)从聚类结果与《词林》的平均一致率来看，高频词明显优于低频词。这是因为词的出现次数越多，统计数据越可靠，因而词义向量的表示越接近真实情况，聚类结果与《词林》平均一致率也就越高。

(3)一般来说，大类不同的义类代码对的聚类结果好于大类相同的义类代码时。如 Ba 06/Da19 比 Aa03/Ae07 好。说明《词林》中义类代码的差距越大，它们对应的词义向量在词义空间中的距离越远，所以聚类结果更容易与《词林》保持一致。

值得强调指出的是：《词林》划分同义词集实现义类编码所依据的是语言学家的直觉或语感，而词义向量的构造则凭借词的同现搭配和大规模语料库的调查，两者的机理完全不同。然而计算机根据词义向量之间的距离对词语实施聚类的

结果却同《词林》的编码高度吻合。这个事实不仅说明李涓子对词义描述提出的两个假设是合理的，而且反映出语言学家的语感在一定程度上也是可以计量的。

综上所述，在《词林》的任意一个同义词集中总有一些词是单义词，而寻找这些单义词在大规模语料库中的同现实词，并分别构造它们的词义向量，完全可以在没有人工干预的条件下自动完成。进而根据假设2，一个同义词集(即一个唯一的义类代码)总可以用一个语义类向量来代表。这个语义类向量是该同义词集中所有单义词的词义向量的质心向量。

语义类向量的每个分量依然是一个同现实词 x_i($i=1$，……，n，n 为向量空间的维数)。设 A 是该语义类向量所代表的同义词集中的所有单义词的集合，则该语义类向量在分量 x_i 上的数值可以用词 x_i 与各个单义词的同现概率 $P(W, x_i)$ 的数学平均值来表征，即

$$V_{xi} = \frac{1}{|A|} \sum_{W \in A} P(w, x_i)$$

其中 $|A|$ 表示单义词集合 A 中词的总数，w 是集合 A 中的任意一个词。

李涓子在一个规模为72兆字节的《人民日报》语料库上，用以上方法为《词林》的每个小类的语义代码分别构造了一个语义类向量，由这些语义类向量形成了一个结构词义空间。在这个词义空间中，词义相近的词所构成的词义向量之间具有较短的距离，从而形成了一个代表这些近义词的语义类向

量;进而距离相近的语义类向量又可以合并成上一层的语义类向量,由此构成了一个如图5-4所示的具有层次结构的词汇词义空间。这项研究的意义在于提出了一种可计算的词义排歧知识表示,这种排歧知识可以利用计算机在大规模语料库中自动获取,因而免除了过去用人工对语料库实施词义标注的巨大开支。

以这种词义排歧的语言模型为基础,对一个多义词的词义辨识过程被简化为以下两步:首先根据这个多义词所在文本(一般来说是一个输入句子)中的上下文形成一个空间向量,然后从词义空间中检索出与该多义词对应的多个语义类向量,它们当中与当前空间向量距离最短的那个语义类向量,就被判定为该多义词当前的词义。

三、基于结构词义空间的词义排歧模型

基于结构词义空间的词义排歧模型如图5-5所示。模型中各模块的功能如下:

(1)特征获取:找出单义词在语料库中的每次出现,把与它前后d($d=7$)个位置上同现的所有实词确定为该词的候选特征。这一步骤只执行一次

(2)特征选择:根据当前多义词的多个义类代码,按候选特征的熵值选出对判别该多义词的词义有表征作用的词,作为该多义词所属各语义类的特征集。比如,“材料”是一个具有三

个义类代码(Ba06/Dk17/Al03)的多义词,特征选择及后继的特征加权都是针对这三个语义类来进行的。所以这是一个动态的排歧过程。

(3)特征加权:选出特征后,计算每个特征对当前多义词所属各语义类的表征能力,作为这些特征的权值。由此形成了针对当前多义词所属各语义类的语义类特征向量。

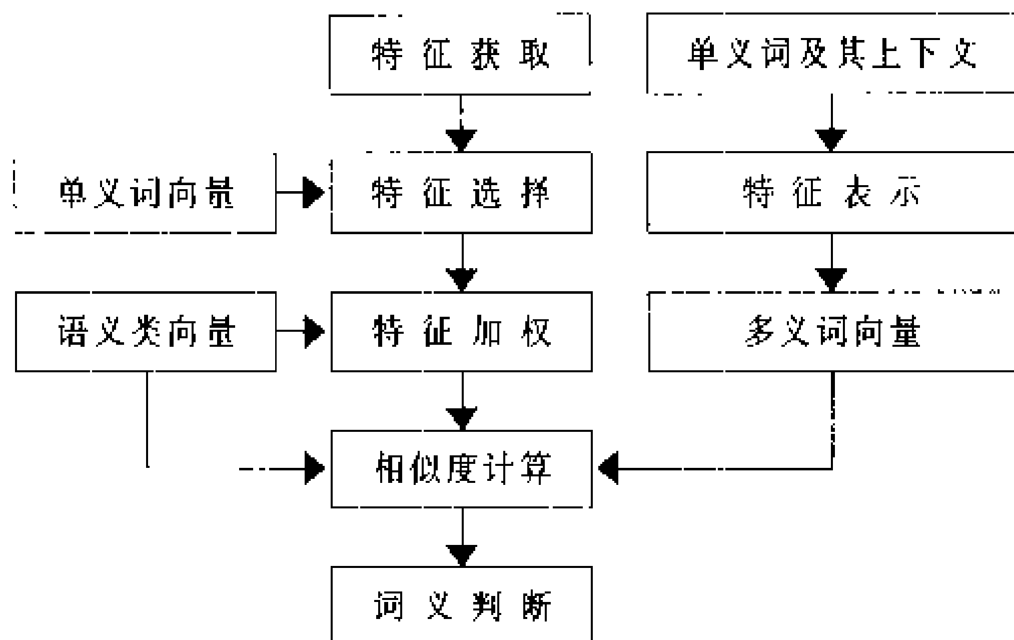


图5-5　基于结构词义空间的词义排歧模型

(4)特征表示:根据以上选出的特征集,把当前多义词所在的上下文用一个特征向量来表示。

(5)相似度计算:分别计算当前多义词的特征向量与其所属各语义类的特征向量之间的相似度。

(6)词义判断:把相似度最高的语义类判定为该多义词当前的词义。如果不存在这样的语义类,则将当前多义词所属的各语义类上升一层,对这个多义词所属上一层的各语义类,重新执行以上排歧过程。

四、词义排歧的实验结果

李涓子进行实验的数据来自《人民日报》新闻语料库，规模为72兆字节。测试时使用的是同类型语料。测试的目的是：1. 检验上文提出的排歧算法是否有效；2. 考察在语料库上用单义词的出现环境构造出的结构词义空间对多义词的词义辨识是否确有帮助。李涓子采用的测试方法分为伪歧义测试和真歧义词测试两种。

(1) 伪歧义词测试

所谓"伪歧义词"是由两个或多个分属不同义类代码的单义词所组成的一个伪"多义词"。针对这里介绍的词义排歧算法，可以把这些单义词进一步限定为具有同一词性。如"收购"和"修改"可组成一个伪歧义词"收购/修改"，其歧义类型为He03/Hg18。

伪歧义测试是一种测试排歧算法的简单方法(Schutze 1992, Gale *et al*. 1992)。它可以节省给测试文本标注词义所耗费的开支。具体做法是首先找出伪歧义所含各词在测试语料中的每次出现，然后用伪歧义词取代它们。这样经过词义排歧后，可以直接利用原先的测试语料来计算词义排歧的正确率。一般来说，采用这种测试方法可以检验一种排歧算法的有效性。

伪歧义测试分为封闭测试和开放测试两类。封闭测试的

测试语料取自训练语料库，从中随机抽取含有每个伪歧义词的 200 个实例。开放测试的测试语料选自训练语料库以外的同类型语料，从中随机抽取含有每个伪歧义词的 100 个实例。词义排歧的正确率定义为：

$$正确率=\frac{词义判断正确的实例数}{测试语料中伪歧义词实例总数}$$

表 5－17 列出 5 个伪歧义词的测试结果。为说明针对伪歧义词所构造的语义类向量的一般性，表 5－18 列出了伪歧义词所含各词及它们所属各次小类在训练料库中的出现次数。

测试结果显示：

a. 对伪歧义词的平均排歧正确率，在封闭测试和开放测试条件下分别达到 93.5% 和 92.6%。说明李涓子提出的基于结构词义空间的词义排歧模型是有效的，她采用的动态特征选择与特征加权方法是合理的。

b. 有些伪歧义词，如"颁布/参与"，虽然在训练语料库中的出现次数较少，但由于它们所属的次小类在训练料库中的出现次数较多，因而仍可获得较高的排歧正确率。可见，利用单义词在语料库上得到的统计数据，基本可以反映语义类的总体分布情况。

(2)真歧义词测试

真歧义词指实际的多义词。测试中有意从不同词类的实词中选出一部分多义词来进行实验。由于这些多义词在构造

语义类向量时肯定没有考察过，所以对真歧义词的测试不存在封闭测试。词义排歧正确率的计算公式仍如前。

表 5－17　伪歧义词测试结果表

伪歧义词	封闭测试正确率	开放测试正确率
权利/事故	97.5%	94.0%
草案/责任	91.0%	89.5%
预算/预赛	98.0%	95.0%
收购/修改	93.0%	93.0%
颁发/参与	92.5%	87.0%
平均正确率	93.5%	92.6%

表 5－18　伪歧义词的歧义情况

伪歧义词	歧义类型	语料库中出现次数	次小类在语料库中出现次数
权利/事故	Di21/Da01	979/959	5088/2187
草案/责任	Dk17/Di22	1563/929	4177/4010
预算/预赛	Hj29/Hh07	841/176	7914/4450
收购/修改	He03/Hg18	954/788	2383/1135
颁布/参与	Hc11/Hi23	449/825	3062/1472

真歧义词的测试结果如表 5－19 所示。实验结果表明，利用单义词在训练语料库上的同现实词来构造结构词义空间的设想是合理的。与同类的无指导词义排歧方法相比，李涓子提出的语言模型具有更高的排歧正确率。更重要的是这种语言模型具备了实行大规模词义排歧的能力，而且这种词义描写方法在原理上也适用于汉语以外的其他自然语言。

表 5-19 真歧义词的开放测试结果表

多义词	歧义类型	测试实例数	次小类在语料库中出现次数	排歧正确率
材料	Dk17/Ba06/Al03	971	1913/1021/422	81.7%
改	Ih02/Hg18/Hj66	2847	1315/1135/309	70.6%
表现	Jd06/Di20/Hj59	754	1323/1500/20	68.9%
发表	Hc11/Hi14/Jd03	2973	5761/2943/214	73.4%
健康	Ed43/Eb37	902	1056/101	70.1%
平均正确率				72.9%

五、结论

(1) 采用基于结构词义空间的词义排歧模型,可以免除对语料库进行词义标注或编制大规模词义排歧知识库的繁重劳动。

(2) 在结构词义空间中,层次越低的语义类向量越能真实地体现该语义类所含同义词的搭配分布情况,因此排歧的正确率越高。建议把这种方法的排歧层次限定为第三层和第四层(即《词林》的小类和次小类)。

(3) 词义排歧结果的优劣与多义词本身的语法特性有关。一般来说,对名词的词义排歧结果好于动词和形容词。对动词来说,带简单宾语的多义词的词义排歧结果优于带复杂宾语(如小句宾语和兼语宾语)的多义词。

(4) 词义排歧结果还与一个多义词的歧义类型有关,多义词所属各语义类之间的距离越小,词义排歧结果越差。

利用大规模语料库构造的结构词义空间，在原理上适合于处理任何多义实词，而且也适用于汉语以外的其他语种。

附录 词性标记集

1. 名词 N
   1.1 普通名词 NG
   1.2 人名 NF
   1.3 地名 NL
   1.4 机关名 NU
2. 时间词 T
3. 处所词 S
4. 方位词 F
5. 动词 V
   5.1 助动词 VA
   5.2 系动词 VI
   5.3 趋向动词 VQ
   5.4 是 VY
   5.5 有 VH
   5.6 名动词 VN
      5.6.1 可带宾语的动词 VNN
      5.6.2 不可带宾语的动词 VNO
   5.7 普通动词 VG
      5.7.1 体宾动词 VGN
      5.7.2 谓宾动词 VGV
      5.7.3 不可带宾语的动词 VGO
6. 形容词 A
7. 状态词 Z
8. 区别词 B
9. 数词 M
   9.1 基数词 MJ
   9.2 序数词 MX
   9.3 其他数词 MG
10. 量词 Q
    10.1 名量词 MQ
    10.2 动量词 QV
11. 代词 R
12. 介词 P
13. 副词 D
14. 连词 C
    14.1 前置连词 CF
    14.2 中置连词 CM
    14.3 后置连词 CN
15. 结构助词 U
    15.1 “的”USDE
    15.2 “地”USDI
    15.3 “得”USDF
    15.4 “似的”USSI
    15.5 “所”USSU
    15.6 “之”USZH
    15.7 时态助词 UT
    15.8 其他助词 UX
16. 语气词 Y
17. 象声词 O

18. 叹词 E

19. 前缀 H

20. 后缀 K

21. 成语 I

22. 简略语 J

23. 习用语 L

24. 其他 X

24.1 非汉字字符串 XCH

24.2 标点符号(各自成一类)

参 考 文 献

[1] 丁信善《语料库语言学的发展及研究现状》,当代语言学,1998,1。

[2] McEnery, T, Wilson. A, Corpus Linguistics, Edinburgh University Press, 1996.

[3] Crystal, D, Stylistic profiling, in Aijmer & Alterberg, 1991, pp. 221 – 238.

[4] Preyer, W, The Mind of a Child, New York: Appletion, 1889.

[5] Stern., Psychology of Early Childhood up to Six Years Old Age, New York: Holt, 1924.

[6] N. Francis, A tagged corpus-problems and prospects, in Greenbaum et al. 1980, pp. 192 – 209. New York: Longman.

[7] G. Kennedy, Preferred ways of putting things, in Svartvik, 1992, pp. 335 – 373.

[8] C. Fries, The Structure of English: An Introduction to the Construction of Sentences, New York: Harcourt-Brace.

[9] Chmosky, N. Syntactic Structures, The Hague: Mouton, 1957.

[10] Halliday, M. A. K. Corpus Studies and Probabilistic Grammar, in Aijmwe and Altenberg, 1991, pp. 30 – 43.

[11] J. Svartvik, Directions in Corpus Linguistics, Berlin: Mouton de

Cruyter, 1992.

[12] Leech, G. (1993),Corpus annotation schemes, Literary and Linguistic Computing. 8(4),pp.275 - 281.

[13] Govindankutty, A., The computer and Dravidian linguistics, ALLC bulletin 1973.

[14] Johansson, S., Continuity and changes in the encoding of computer corpora, in Oostdijk & Hann,1994,pp.13 - 31.

[15] Kennedy, G., An Introduction to Corpus Linguistic, London & New York: Longdon, 1998.

[16] Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J.(1985) A Comprehensive Grammar of the English Language, London: Longman.

[17] Sinclair, J. (1991) The automatic analysis of corpora, In Svartvik 1992, pp.379 - 397.

[18] G. Leech, (1991) "The State of Art in Corpus Linguistics", in Aijmer and Altenberg 1991, pp.8 - 29.

[19] Summers, D. Longman/Lancaster English Language Corpus: Criteria and Design, Harlow: Longman,1991.

[20] 张普《关于大规模真实文本语料库的几点理论考虑》,语言文字与应用,1999,1。

[21] Runddle, M.,Stock. P., The corpus revolution, English Today 1992.

[22] Biber, D. 1993, 'Representativeness in Corpus Design', Literary and Linguistic Computing vol.8,No.4;pp.243 - 257.

[23] Sinclair, J. Corpus, Concordence, Collocation. Oxford University, 1991.

[24] 马少平《脱机手写体汉字识别方法与系统》,博士论文,清华大学计算机科学与技术系,1997,5。

[25] 国家语言文字工作委员会汉字处《现代汉语常用字表》,语文出版

社。

[26] Walker, D. E., The Ecology of Language, in the Proceedsings of International Workshop on Electronic Dictionaries, 1990, Oiso Japan.

[27] Quirk, R., Towards a Description of English Usage, Transaction of Philological Society, pp.40 – 61.

[28] Fransic, W. N., Kucera, H. Frequency Analysis of English Usage: Lexicon and Grammar Boston: Houghton Mifflin.

[29] Greene, B. B. Rubin, G. M. Automatic grammatical tagging of English, Providence, R. I.: Department of Linguistics, Brown University.

[30] Brown Corpus: http://khnt.hit.uib.no/icame/manuals/brown/index.htm.

[31] Johansson, S., Atwell, E., Carside, R. And Leech, G. 1986. The tagged LOB corpus: User's Manual, Norwegian Computing Centre for the Humanities, Bergen.

[32] Jonansson, S. (ed.) Computer Corpora in English Language Research, Bergen: Norwegian Computing Centre for the Humanities. 1982.

[33] Johansson, S., Hofland, L., Frequency Analysis of English Vocabulary and Grammar 2 vols. Oxford: Clarendon Press.

[34] Svartvik, J. (ed.) The London-Lund Corpus of Spoken English: Description and Research Lund: Lund Studies in English 82. Lund University Press.

[35] The Bank of English, http://titania.cobuild.collins.co.uk/boe-info.html.

[36] Collins COBUILD English Dictionary, Collins COBUILD.

[37] Della Summers, Longman Group UK, Longman/Lancaster English Language Corpus-Criteria and Design, International Journal of

Lexicography, Vol. 6 No. 3: pp.181 - 208.
[38] British National Corpus, http://info.ox.ac.uk./bnc.
[39] The International Corpus of English http: //www. ucl. ac. uk/ Enghish-usage/ice. htm.
[40] 陈鹤琴《语体文应用字汇》,商务印书馆,1928。
[41] 王还、常宝儒等《现代汉语频率词典》,北京语言学院出版社, 1986。
[42] 陈原《现代汉语定量分析》,上海教育出版社,1989。
[43] 台湾中央研究院平衡语料库,http://godel.iis.sinica.edu.tw
[44] Huang, Chu-Ren and Keh-Jiann Chen, A Chinese Corpus for Linguistic Research. Proceedings of the 1992 International Conference on Computational Linguistics (Nantes, France, 1992), pp.1214 - 1217.
[45] Huang, Chu-Ren and Keh-Jiann Chen, Modern and Classical Chinese Corpora at Academic Sinica Text Databases for Natural Language Processing and Linguistic Computing, Presented at the Sixth CODATA Task Group Meeting on the Survey of Data Sources in Asian-Oceanic Countries (Taipei: Academic Sinica, 1994).
[46] K. J. Chen, C. R. Huang, L.P. Chang, H. L. Hsu, 1996, "SINICA CORPUS: Design Methodology for Balanced Corpora," Proceedings of PACLICLL, pp.167 - 176, Seoul, Korea.
[47] 香港城市大学语言资讯科学研究中心《中文各地区共时词语研究报告》,1998,5。
[48] 孙宏林、黄建平等《现代汉语语料库系统鉴定会文件》,1996,1。
[49] 王建新《索引软件:语料库语言学的有利工具》,当代语言学, 1998,1。
[50] Biber, D., Finegan, E., Intra-textual variation within medical research articles, in Oostdijk & de Haan, 1994,pp.201 - 222.

[51] Kjellmer, G. (1991)'A mint of phrases', in Aijmer and Altenberg 1991, pp.111-127.

[52] 朱雪龙、艾红梅《应用信息论基础》,清华大学内部教材,1998。

[53] Leech, G. (1992), Corpora and theories of linguistic performance, in J. Svartvik 1992, pp.149-163.

[54] Lari K. Young S. Applications of stochastic context-free grammars using the inside-outside algorithm, In: Computer Speech & Language 1991,5, pp.237-257.

[55] Pereira, F., Schabes, Y. Inside-outside reestimation from partially bracket corpora, In: Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, University of Delaware, Newark, Delaware, USA, 1995, pp.128-135.

[56] Eugence Charniak, Statistical Language Learning, The MIT Press, London, England.

[57] John Sinclair, Language independent statistical software for corpus exploration. Computers and the Humanities, 1998, Vol.31, pp.229-255.

[58] Garside, R., Leech, G., McEnery, A., Corpus Annotation, Longman & New York, 1997.

[59] 刘开瑛《中文文本自动分词和标注》,北京:商务印书馆,2000。

[60] Ellegard, A.(1978)The syntactic structure of English texts: A computer-based study of four kinds of text in Brown University Corpus. Guöteborg:Gothenburg Studies in English 43.

[61] Van Halteren, H., Oostdijk, N., Towards a syntactic database: The TOSGA analysis system. In Arts et al. 1993, pp.145-162.

[62] Leech, G. N., Garside, R., Running a grammar factory: the production of syntactically analysed corpora or "treebanks", In Johansson and Stenstron, 1996, pp.15-32.

[63] The Penn Treebank Project, http://www.cis.upenn.edu/~treebank/home.html.

[64] Chelba, C., Exploiting Syntactic Language Structural for Language Modeling, A Dissertation of John Hopskin University, January, 2000.

[65] Collins, M. J., A New Statistical Parser Based on Lexical Dependencies, The proceedings of the 34th Annual Meeting of the ACL, Santa Cruz, California 1996.

[66] Jelinek, F., Lafferty, J. And Mercer, R., *et al*, Decision Tree Parsing Using a Hidden Derivation Model, the proceedings of the 1994 Human Language Technology Workshop, pp.272-277.

[67] Richardson, S. D., Dolan, W. B., Vandewende, L., MindNet: Acquring and Structuring Semantic Information from Text, ACL'98, vol. 2, pp.1098-1102.

[68] Church, K., Gale, W., Hanks, P. And Hindle, D., Using statistic in lexical analysis, in Zernik 1991, pp.115-164.

[69] Schmidt, K. M. Qualitative and quantitative research approaches to English constructions, in Souter and Atwell 1993, pp.85-96.

[70] Mindt, D. Syntactic evidence for semantic distinctions in English, Aijmer and Altenberg 1991, pp.182-196.

[71] Stenstrom, A. B. Carry on signals in English conversation, in Meijs 1987, pp.87-119.

[72] 赵淑华《现代汉语句型统计与研究》,成果报告,北京语言文化大学,1995,4,10。

[73] AHI 语料库:http://ref.umdl.umich.edu/a/ahd/sample.htm

[74] Carroll, J.B., Davies, P.And Richman, B. The American Heritage Word Frequency Book, New York: American Heritage Publishing Co.

[75] Collins Cobuild English Language Dictionary, 1987, London,

Collins.

[76] 黄居仁、陈克健等《国语日报量词典》,台北:国语日报社,1997.

[77] Firth, J. R.. A synopsis of linguistic theory, 1930 - 1955, in Studies in Linguistic Analysis Oxford: Blackwell.

[78] Essex. (new ed) Longman dictionary of contemporary English, England: Longman.

[79] Summers, D. Longman Language Activator, 1993, Longman.

[80] Biber, D., Finegan, E., On the exploitation of computerized corpora in variation studies, in Aijmer & Altenberg,1991,pp.204 - 220.

[81] Benson, M., Benson, E. And Ilson, R., The BBI Combinatory Dictionary of English, Amsterdam: John Benjamins Publishing Co., 1986.

[82] Benson, M., A Combinatory Dictionary of English, Dictionaries: Journal of the Dictionary Society of North America, 7.

[83] 张寿康、林杏光《现代汉语实词搭配词典》、北京:商务印书馆, 1992。

[84] Choueka, Y., Klein, T., and Neuwitz, E., Automatic retrieval of frequent idiomatic and collocation expressions in a large corpus, Journal of Literary and Linguistic Computing, 4.

[85] Church, K., Hanks, P., Word Association, Mutual Information and Lexicography. In Proceedings of 27^{th} Annual Meeting of Association for Computational Linguistics, 1989, pp.76 - 83.

[86] Smadja, F., Retrieving Collocation from Text: Xtract, Computational Linguistics, vol.19,No.1.

[87] 孙茂松等《汉语搭配定量分析初探》,中国语文, 1997,1。

[88] Ahrens, Kathleen and Chu-Ren Huang, Classifiers and Semantic Type Coercion: Motivating a New Classification of Classifiers, In. B.-S. Park and J.B. Kim. eds. Proceeding of the 11th Pa-

cific Asia Conference on Language, Information and Computation (Seoul: Kyung Hee University, 1996), pp.1 10.

[89] Shannon, C. A mathematical theory of communications, Bell System Technical Journal, 1949.27, pp.623－656.

[90] Bible, D., Conrad, S. And Reppen, R., Corpus-based Approaches to issues in applied linguisitics, Applied Linguistics vol. 15, No.2, pp.169－189.

[91] Biber, D., Variation across Speech and Writing, Cambridge: Cambridge University Press, 1988.

[92] 刘开瑛《自动分词与词性标注评测》,计算机世界,评测专版,1996,3,25。

[93] 孙茂松、黄昌宁等《零用汉字二元语法关系解决汉语自动分词中交集型歧义》,计算机研究与发展,1997年,第34卷第5期。

[94] 吴芳芳《自动分词中歧义字段切分方法研究》,硕士论文,山西大学,1998.

[95] 左正平《汉语自动分词中的若干问题》,清华大学计算机科学与技术系硕士论文,1998,6。

[96] 孙茂松、左正平《汉语真实文本中的交集型歧义》,汉语计量与计算研究,1998。

[97] Church, K., A stochastic parts program and noun phrase parser for unrestricted text, In: Proceedings of the Second Conference on Applied Natural Language Processing, 1988.

[98] 李文捷、潘海华等《基于语料库的中文最长名词短语的自动抽取》;陈力为、袁琦编:《计算语言学进展与应用》,北京:清华大学出版社,1995,pp.119－125。

[99] 赵军《汉语基本名词短语的识别和结构分析研究》,博士论文,清华大学计算机科学和技术系,1998。

[100] 张卫国《三种定语、三类意义及三个槽位》,中国人民大学学报,1996,No.4, pp.97－100。

[101] Brill, E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging, In: Computational Linguistics, V21. No.4, 1995.

[102] Ramshaw, L., Marcus R. Text chunking using transformation-based learning, In: Proceedings of the Fourth Workshop on Very Large Corpus, 1995, pp.82~94.

[103] Lesk, Michael, Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the 1986 SIGDOC conference, pp.24-26.

[104] Wilks, Yorick A. and Dan Fass. Preference semantics: A family histroy. Report MCCS-90-194, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

[105] Yarowsky David, Word sense disamhiguation using statistical model of Roget's categories trained on large corpora. Proceedings of the 14th International Conference on Computational Linguistics, COLING'92, pp.454-460, Nantes, France, 1992, August

[106] D. Yarowsky, Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In: Proc 32nd Annual Meeting of Association for Computational Linguistics, 1994, pp.88-95.

[107] Bruce R. A Statistical Method for Word Sense Disambiguation, [Ph. D. Dissertation] USA: New Mexico State University, 1995, pp.43 78.

[108] 梅家驹、竺一鸣、高蕴琦《同义词词林》,上海:上海辞书出版社,1983。

[109] Firth, J.R. 1957. Modes of Meaning. Papers in Linguistics 1934-1951, pp.190-215, Oxford University Press, Oxford, UK.